(19)日本国特許庁（ＪＰ）　(12)**公表特許公報**（Ａ）　(11)特許出願公表番号
特表2001－519070
（P2001－519070A）
(43)公表日　平成13年10月16日(2001.10.16)

(51)Int.Cl.⁷	識別記号	Ｆ Ｉ	テーマコード（参考）
Ｇ０６Ｆ　17/30	１７０	Ｇ０６Ｆ　17/30	１７０Ｆ
Ａ６１Ｋ　38/00		Ａ６１Ｋ　39/21	
39/21		Ａ６１Ｐ　31/18	
Ａ６１Ｐ　31/18		Ｃ０７Ｋ　14/155	
Ｃ０７Ｋ　14/155		Ｇ０１Ｎ　33/68	

審査請求　未請求　予備審査請求　有　（全186頁）　最終頁に続く

(21)出願番号　特願平10－544599
(86)（22)出願日　平成10年３月23日(1998.3.23)
(85)翻訳文提出日　平成11年９月24日(1999.9.24)
(86)国際出願番号　ＰＣＴ／ＣＡ９８／００２７３
(87)国際公開番号　ＷＯ９８／４３１８２
(87)国際公開日　平成10年10月１日(1998.10.1)
(31)優先権主張番号　６０／０４１，４７２
(32)優先日　平成９年３月24日(1997.3.24)
(33)優先権主張国　米国（ＵＳ）

(71)出願人　クイーンズ　ユニバーシティー　アット　キングストン
カナダ国　オンタリオ州　キングストン　（番地なし）
(72)発明者　スティーグ　エヴァン　ダブリュー.
カナダ国　オンタリオ州　キングストン　オンタリオストリート　32　アパートメント　709
(74)代理人　弁理士　清水　初志　（外１名）

最終頁に続く

(54)【発明の名称】　一致検出の方法、製品および装置

(57)【要約】
各対象が多数の属性を有する、対象のデータセットにおける一致を検出するための方法およびシステムについて開示する。データセットの等しい大きさの部分集合が反復的にサンプリングされ、一致（部分集合内の1つまたは複数の対象における複数の属性の値の同時出現）が記録される。関心対象の各一致に関して、期待される一致の数が決定され、観測された一致数と比較される。この比較は一致に関して複数の属性の相関の程度を決定するために用いられる。その結果として得られた、相関の程度があらかじめ決定された閾値を上回る複数の属性である、k項数の相関属性の集合が報告される。本方法およびシステム（プロセッシングノードのアレイ上に実装される）は、HIV研究などにおける蛋白質構造解析に適する。

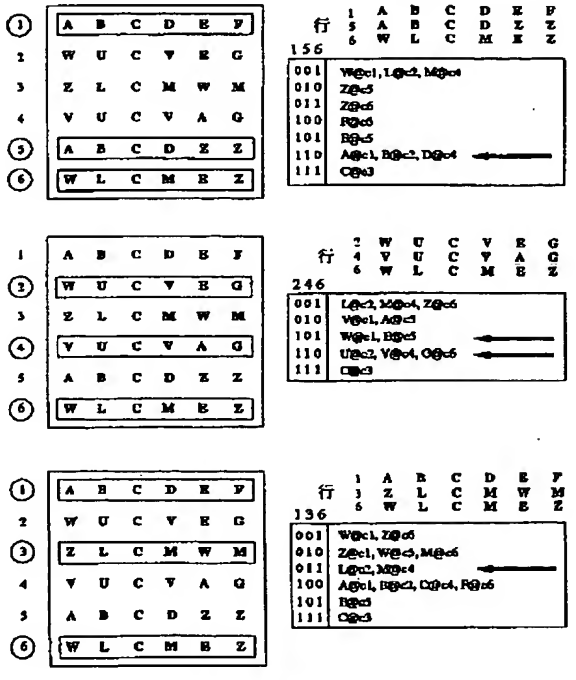


図5

【特許請求の範囲】

１．多数の属性を有するデータセットとともに用いるための一致検出法であって、下記の段階を含む方法：

・対象がある属性を有する場合にその属性が対象において出現すると呼ばれる、N_A個の変数（「属性」）を項とするM個の対象の集合の表示、

・あらかじめ決定された数の反復における各反復に関する、M個の対象からのr_i個の部分集合のサンプリング、

・一致がサンプリングされた部分集合におけるr_i個の対象のうち同じh_i個における$1 \leqq k \leqq N_A$個の属性の同時出現であって、$0 \leqq h_i \leqq r_i$である、サンプリングされた対象の部分集合のそれぞれにおけるk個の属性の集合間での一致の検出および記録、

・決定がサンプリングおよび収集の前、同時またはサンプリングおよび収集の後に行われる、上記のk個の属性の任意の集合ならびにあらかじめ決定された数のサンプリングおよび一致計数の反復に関する一致の期待数の決定、

・k個の属性の任意の集合ならびにサンプリングおよび一致計数の反復回数に関する一致の観測値と期待数との比較、ならびにこの比較によるk個の属性の集合に関する相関（または結合もしくは依存）の程度の決定、ならびに

・k項数の相関属性が、選択された相関の程度に関してあらかじめ決定された閾値を上回る値を持つことがこの過程によって決定されたN_A個の属性のkの集合である、k項数の相関属性の集合の報告。

２．多数の属性を有する対象のデータセットとともに用いるための一致検出法であって、下記の段階を含む方法：

・各反復でデータセットのサンプリングされた部分集合が各対象に関して属性の同じ部分集合を有する、あらかじめ決定された数の反復にわたるデータセットの部分集合のサンプリング、

・一致がデータセットのサンプリングされた部分集合における1つまたは複数の対象における複数の属性値の同時出現であって、複数の属性値が各出現に関して同一であって、データセットのサンプリングされた各部分集合における一致の検出および数の記録が他の部分集合における一致のサンプリング、検出および数の

記

録の前、同時または後に行われる、データセットのサンプリングされた各部分集合における一致の検出および数の記録、

・サンプリング、検出および記録の前、同時または後に行われる、関心対象の各一致に関する期待数の決定、

・関心対象の各一致に関する一致の観測数と一致の期待数との比較、およびこの比較による、一致に関する複数の属性の相関の程度の決定、ならびに

・k項数の相関属性が、それに関する相関の程度がそれぞれにあらかじめ決定された閾値を上回る複数の属性である、k項数の相関属性の集合の報告。

３．観測数と期待数との比較が尾部確率に関するチャーノフ境界を用いて計算される、請求項2記載の一致検出法。

４．数がサンプリングされた部分集合のすべてにわたって各一致の数の連続的な合計を保存することによって記録される、請求項2記載の一致検出法。

５．多数の属性を有する対象のデータセットの視覚的表示のための方法であって、下記の段階を含む方法：

・各反復でデータセットのサンプリングされた部分集合が各対象に関して属性の同じ部分集合を有する、あらかじめ決定された数の反復にわたるデータセットの部分集合のサンプリング、

・一致がデータセットのサンプリングされた部分集合における1つまたは複数の対象における複数の属性値の同時出現であって、複数の属性値が各出現に関して同一であって、データセットのサンプリングされた各部分集合における一致の検出および数の記録が他の部分集合における一致のサンプリング、検出および数の記録の前、同時または後に行われる、データセットのサンプリングされた各部分集合における一致の検出および数の記録、

・サンプリング、検出および記録の前、同時または後に行われる、関心対象の各一致に関する期待数の決定、

・関心対象の各一致に関する一致の観測数と一致の期待数との比較、およびこの比較による、一致に関する複数の属性の相関の程度の決定、ならびに

・k項数の相関属性が、それに関する相関の程度がそれぞれにあらかじめ決定された閾値を上回る複数の属性である、k項数の相関属性の集合の、グラフィカルイン

ターフェースを通じてのユーザーに対する報告。

６．多数の属性を有する対象のデータセットの高次相互作用を捕捉してデータモデル化ユニットに報告するための、データモデル化ユニットとともに用いるための予備処理の方法であって、下記の段階を含む方法：

・各反復でデータセットのサンプリングされた部分集合が各対象に関して属性の同じ部分集合を有する、あらかじめ決定された数の反復にわたるデータセットの部分集合のサンプリング、

・一致がデータセットのサンプリングされた部分集合における1つまたは複数の対象における複数の属性値の同時出現であって、複数の属性値が各出現に関して同一であって、データセットのサンプリングされた各部分集合における一致の検出および数の記録が他の部分集合における一致のサンプリング、検出および数の記録の前、同時または後に行われる、データセットのサンプリングされた各部分集合における一致の検出および数の記録、

・サンプリング、検出および記録の前、同時または後に行われる、関心対象の各一致に関する期待数の決定、

・関心対象の各一致に関する一致の観測数と一致の期待数との比較、およびこの比較による、一致に関する複数の属性の相関の程度の決定、ならびに

・k項数の相関属性が、それに関する相関の程度がそれぞれにあらかじめ決定された閾値を上回る複数の属性である、k項数の相関属性の集合の、データモデル化ユニットへの報告。

７．多数の属性を有する対象のデータセットとともに用いるための相関消去の方法であって、下記の段階を含む方法：

・各反復でデータセットのサンプリングされた部分集合が各対象に関して属性の同じ部分集合を有する、あらかじめ決定された数の反復にわたるデータセットの部分集合のサンプリング、

・一致がデータセットのサンプリングされた部分集合における1つまたは複数の対象における複数の属性値の同時出現であって、複数の属性値が各出現に関して同一であって、データセットのサンプリングされた各部分集合における一致の検出および数の記録が他の部分集合における一致のサンプリング、検出および数の記

録の前、同時または後に行われる、データセットのサンプリングされた各部分集合における一致の検出および数の記録、

・サンプリング、検出および記録の前、同時または後に行われる、関心対象の各一致に関する期待数の決定、

・関心対象の各一致に関する一致の観測数と一致の期待数との比較、およびこの比較による、一致に関する複数の属性の相関の程度の決定、ならびに

・k項数の相関属性が、それに関する相関の程度がそれぞれにあらかじめ決定された閾値を上回る複数の属性である、k項数の相関属性の集合の消去。

８．対象が、各トランザクションが1つまたは複数の購入された製品を含む販売トランザクションであって、属性が特定の製品または特定の種類の製品の販売の事例である、請求項2記載の方法。

９．対象が時間刻みであって属性がシステムにおける要素の状態である、請求項2記載の方法。

10．対象が時間刻みであって属性が金融証書または商品の価格または価格の変動である、請求項2記載の方法。

11．方法の段階が以下の疑似コードによって表現される請求項２記載の方法：

```
0. begin
1. read(MATRIX);
2. read(R, T);
3. compute_first_order_marginals(MATRIX);
4. csets:={};
5. for iter=1 to T do
6. sampled_rows:=rsample(R, MATRIX):
```

```
7. attributes:=get_attributes(sampled_rows);
8. all_coincidences:=find_all_coincidences(attributes);
9. for coincidence in all_coincidences do
10. if cset_already_exists(coincidence,csets)
11. then update_cset(coincidence, csets);
12. else add_new_cset(coincidence,csets);
13. endif
14. endfor
15. endfor
16. for cset in csets do
17. expected:=compute_expected_match_count(cset);
18. observed:=get_observed_match_count(cset);
19. stats:=update_stats(cset,hypoth_test(expected,observed));
20. endfor
21. print_final_stats(csets,stats);
22. end。
```

12. 各対象が多数の属性を有する、対象のデータセットとともに用いるための一致検出システムであって、下記の手段を含むシステム：

・各反復でデータセットのサンプリングされた部分集合が各対象に関して属性の同じ部分集合を有する、あらかじめ決定された数の反復にわたるデータセットの部分集合のサンプリングのための手段、

・一致がデータセットのサンプリングされた部分集合における1つまたは複数の対象における複数の属性値の同時出現であって、複数の属性値が各出現に関して同一であって、データセットのサンプリングされた各部分集合における一致の検出および数の記録が他の部分集合における一致のサンプリング、検出および数の記録の前、同時または後に行われる、データセットのサンプリングされた各部分集合における一致の検出および数の記録のための手段、

・サンプリング、検出および記録の前、同時または後に行われる、関心対象の各

一致に関する期待数の決定のための手段、

・関心対象の各一致に関する一致の観測数と一致の期待数との比較、およびこの比較による、一致に関する複数の属性の相関の程度の決定のための手段、ならびに

・k項数の相関属性が、それに関する相関の程度がそれぞれにあらかじめ決定された閾値を上回る複数の属性である、k項数の相関属性の集合の報告のための手段。

13. 集合体（aggregate）におけるシステムの手段が以下の疑似コードによって表

現される方法を実施する、請求項１２記載の一致検出システム：

```
0. begin
1. read(MATRIX);
2. read(R, T);
3. compute_first_order_marginals(MATRIX);
4. csets:={};
5. for iter=1 to T do
6. sampled_rows:=rsample(R, MATRIX):
7. attributes:=get_attributes(sampled_rows);
8. all_coincidences:=find_all_coincidences(attributes);
9. for_coincidence in all_coincidences do
10. if cset_already-exists(coincidence,csets)
11. the nupdate_cset(coincidence, csets);
12. else add_new_cset(coincidence, csets);
13. endif
14. endfor
15. endfor
16. for cset in csets do
17. expected:=compute_expected_match_count(cset);
```

18. observed:=get_observed_match_count(cset);

19. stats:=update_stats(cset, hypoth_test(expected, observed));

20. endfor

21. print_final_stats(csets, stats);

22. end。

14. データセットの部分集合のサンプリングのための手段がデータセットを分割してサンプリング用の部分集合にするための手段を含む、請求項12記載の一致検出システム。

15. 一致の検出および数の記録のための手段が、それぞれのプロセッシングノードが一致の検出および各サブカウントの記録をするプロセッシングノードのアレイを含むことができ、関心対象の各一致に関して該一致の観測数と前記一致の期待数とを比較するための手段が、該サブカウントをマージして該観測数を提供するための手段を含む、請求項14記載の一致検出システム。

16. プロセッシングノードの少なくとも1つがそれぞれの一致のサブサブカウントを検出および記録するそれぞれのプロセッシングノードのサブアレイを含み、マージのための手段が該サブサブカウントをマージしてサブカウントおよび／または観測数を提供する、請求項15記載の一致検出システム。

17. 各プロセッシングソードが、データセットの受け取られた部分集合を保存するための入力バッファーおよびサブカウントまたはサブサブカウントを保存するための出力バッファーを含むメモリ、ならびにメモリとの間でデータをやり取りするメモリバスを含む、請求項15または16記載の一致検出システム。

18. コンピュータと、対象対属性の行列の形式で表現された多数の属性を有する対象のデータセットと共に用いるための一致検出プログラム媒体であって、

そのコンピュータと互換性がある保存媒体上に保存されたコンピュータプログラムであって、

・各反復でデータセットのサンプリングされた部分集合が各対象に関して属性の同じ部分集合を有する、あらかじめ決定された数の反復にわたるデータセットの部分集合のサンプリング、

・一致がデータセットのサンプリングされた部分集合における1つまたは複数の対象における複数の属性値の同時出現であって、複数の属性値が各出現に関して同一であって、データセットのサンプリングされた各部分集合における一致の検出および数の記録が他の部分集合における一致のサンプリング、検出および数の記録の前、同時または後に行われる、データセットのサンプリングされた各部分集合における一致の検出および数の記録、

・サンプリング、検出および記録の前、同時または後に行われる、関心対象の各一致に関する期待数の決定、

・関心対象の各一致に関する一致の観測数と一致の期待数との比較、およびこの比較による、一致に関する複数の属性の相関の程度の決定、ならびに

・k項数の相関属性が、それに関する相関の程度がそれぞれにあらかじめ決定さ

れた閾値を上回る複数の属性である、k項数の相関属性の報告

のためにコンピュータを指向させる指示を含むコンピュータプログラムを含む媒体。

19. 多数の属性を有する対象のデータセットとともに用いるための一致検出システムであって、

コンピュータ、ならびに

そのコンピュータと互換性のある媒体上のコンピュータプログラムであって、

・各反復でデータセットのサンプリングされた部分集合が各対象に関して属性の同じ部分集合を有する、あらかじめ決定された数の反復にわたるデータセットの部分集合のサンプリング、

・一致がデータセットのサンプリングされた部分集合における1つまたは複数の対象における複数の属性値の同時出現であって、複数の属性値が各出現に関して同一であって、データセットのサンプリングされた各部分集合における一致の検出および数の記録が他の部分集合における一致のサンプリング、検出および数の記録の前、同時または後に行われる、データセットのサンプリングされた各部分集合における一致の検出および数の記録、

・サンプリング、検出および記録の前、同時または後に行われる、関心対象の各一致に関する期待数の決定、

・関心対象の各一致に関する一致の観測数と一致の期待数との比較、およびこの比較による、一致に関する複数の属性の相関の程度の決定、ならびに

・k項数の相関属性が、それに関する相関の程度がそれぞれにあらかじめ決定された閾値を上回る複数の属性である、k項数の相関属性の集合の報告

のためにコンピュータを指向させるコンピュータプログラム

を含むシステム。

20. データセットのサンプリングの前に、データセットがその行列のサンプリングによってサンプリングされる、対象と属性からなる行列の形で対象および属性を提示する段階をさらに含む、請求項2記載の一致検出法。

21. ・各反復でデータセットのサンプリングされた部分集合が、各対象に関して属性の同じ部分集合を有する、あらかじめ決定された数の反復に関する対象対属

性を表現するデータセットの部分集合のサンプリング、

・一致がデータセットのサンプリングされた部分集合における1つまたは複数の対象における複数の属性値の同時出現であって、複数の属性値が各出現に関して同一であって、データセットのサンプリングされた各部分集合における一致の検出および数の記録が他の部分集合における一致のサンプリング、検出および数の記録の前、同時または後に行われる、データセットのサンプリングされた各部分集合における一致の検出および数の記録、

・サンプリング、検出および記録の前、同時または後に行われる、関心対象の各一致に関する期待数の決定、

・関心対象の各一致に関する一致の観測数と一致の期待数との比較、およびこの比較による、一致に関する複数の属性の相関の程度の決定、ならびに

・k項数の相関属性が、それに関する相関の程度がそれぞれにあらかじめ決定された閾値を上回る複数の属性である、k項数の相関属性の集合の報告

によって選択される属性の集合を有する製品。

22. ・各反復でデータセットのサンプリングされた部分集合が各対象に関して属

性の同じ部分集合を有する、あらかじめ決定された数の反復に関する対象対属性を表現するデータセットの部分集合のサンプリング、

・一致がデータセットのサンプリングされた部分集合における1つまたは複数の対象における複数の属性値の同時出現であって、複数の属性値が各出現に関して同一であって、データセットのサンプリングされた各部分集合における一致の検出および数の記録が他の部分集合における一致のサンプリング、検出および数の記録の前、同時または後に行われる、データセットのサンプリングされた各部分集合における一致の検出および数の記録、

・サンプリング、検出および記録の前、同時または後に行われる、関心対象の各一致に関する期待数の決定、

・関心対象の各一致に関する一致の観測数と一致の期待数との比較、およびこの比較による、一致に関する複数の属性の相関の程度の決定、ならびに

・k項数の相関属性が、それに関する相関の程度がそれぞれにあらかじめ決定された閾値を上回る複数の属性である、k項数の相関属性の集合の報告

によって生成される一組の規則を適用するこによって規定される製品。

23. ・請求項2記載の方法、および

・報告された相関属性によって規定される規則の適用

の段階をさらに含む方法。

24. 残基A18/Q31/H33の空間座標を含むHIVエンベロープ蛋白質のV3ループの構造モチーフを含むペプチドまたは疑似ペプチド。

25. 請求項2の方法を用いて同定される構造モチーフを有する蛋白質と相互作用するリガンドと、その薬学的に許容される担体または賦形剤とを含む薬学的組成物。

26. リガンドが、適した実体を有し且つその成分が対応するモチーフの残基または部分と相互作用するような位置に互いにある化学的成分を含む、請求項25記載の薬学的組成物。

27. リガンドがモチーフとの相互作用によってモチーフを含む蛋白質の領域の機能を妨げる、請求項26記載の薬学的組成物。

28. 請求項2記載の方法を用いて同定される構造モチーフを有する蛋白質と相互作用するリガンド、およびそのリガンドと結合した検出可能な標識を含む診断薬。

29. 残基A18/Q31/H33の空間座標を有するV3ループの構造モチーフを含むエンベロープ蛋白質であって、該モチーフと相互作用する官能基を少なくとも1つ含むリガンドとその薬学的に許容される担体または賦形剤とを含む、ヒト免疫不全ウイルス（HIV）のエンベロープ蛋白質と相互作用する薬学的組成物。

30. リガンドが、残基18との結合能を有していて前記リガンド中の残基18との結合のための有効部分に存在する少なくとも1つの官能基、残基31との結合能を有していて前記リガンド中の残基31との結合のための有効部分に存在する少なくとも1つの官能基、および残基33との結合能を有していて前記リガンド中の残基33との結合のための有効部分に存在する少なくとも1つの官能基を含む、請求項29記載の薬学的組成物。

31. ヒト免疫不全ウイルス（HIV）のエンベロープ蛋白質の構造モチーフと相互作用するリガンドを設計する方法であって、HIVエンベロープ蛋白質のV3ループ内の残基A18、Q31およびH33の空間座標を有するテンプレートの提供、ならびに空間的

拘束を有する有効なアルゴリズムを用いての化学的リガンドの計算的な展開（evolving）であって該展開されたリガンドがモチーフと結合する官能基を少なくとも1つ含むような展開の段階を含む方法。

32. リガンドが、残基18との結合能を有していて前記リガンド中の残基18との結合のための有効部分に存在する少なくとも1つの官能基、残基31との結合能を有していて前記リガンド中の残基31との結合のための有効部分に存在する少なくとも1つの官能基、および残基33との結合能を有していて前記リガンド中の残基33との結合のための有効部分に存在する少なくとも1つの官能基を含む、請求項31記載の方法。

33. ヒト免疫不全ウイルス（HIV）のエンベロープ蛋白質の構造モチーフと結合するリガンドを同定する方法であって、HIVエンベロープ蛋白質のV3ループ内の

残基A18、Q31およびH33の空間座標を有するテンプレートの提供、分子の構造および配向性を含むデータベースの提供、ならびにその成分がモチーフと相互作用するように互いに対して配置された有効成分を該分子が含むかどうかを決定するための該分子のスクリーニングの段階を含む方法。

34．分子の第1の成分が残基18と相互作用し、分子の第2の成分が残基31と相互作用し、分子の第3の成分が残基33と相互作用する、請求項33記載の方法。

35．本明細書に記載される共変するk項数を具現化した抗原およびワクチン。

36．・各反復でデータセットのサンプリングされた部分集合が各対象に関して属性の同じ部分集合を有する、あらかじめ決定された数の反復に関する対象対属性を表現するデータセットの部分集合のサンプリング、

・一致がデータセットのサンプリングされた部分集合における1つまたは複数の対象における複数の属性値の同時出現であって、複数の属性値が各出現に関して同一であって、データセットのサンプリングされた各部分集合における一致の検出および数の記録が他の部分集合における一致のサンプリング、検出および数の記録の前、同時または後に行われる、データセットのサンプリングされた各部分集合における一致の検出および数の記録、

・サンプリング、検出および記録の前、同時または後に行われる、関心対象の各一致に関する期待数の決定、

・関心対象の各一致に関する一致の観測数と一致の期待数との比較、およびこの比較による、一致に関する複数の属性の相関の程度の決定、ならびに

・k項数の相関属性が、それに関する相関の程度がそれぞれにあらかじめ決定された閾値を上回る複数の属性である、k項数の相関属性の集合の報告

によって選択される一組の属性との相互作用によって規定される製品。

37．対象が化合物であって属性が特定の化学成分を含む、請求項2記載の方法。

38．対象がペプチドまたは蛋白質であって属性がモチーフの特定の構造または下部構造のパターンを含む、請求項2記載の方法。

39．対象が化合物、分子構造、ヌクレオチド配列およびアミノ酸配列からなる群より選択され、属性が選択された対象の特徴である、請求項2記載の方法。

40. 対象が時間刻みであって属性が遺伝子または遺伝子産物の生物的パラメータである、請求項2記載の方法。

41. 対象が電子的に保存されるおよび／または電子的に索引が付けられた（indexed）文書であって属性が題目である、請求項2記載の方法。

42. 対象が消費者であって属性が該消費者によって購入された、または購入されなかった製品を含む、請求項2記載の方法。

43. 属性が、消費者に対して郵送されたこと、またはされなかったことをさらに含む、請求項42記載の方法。

44. 対象が製品を含み、属性がそれらの製品を購入した、または購入しなかった消費者を含む、請求項2記載の方法。

45. 属性が消費者の人口統計変数をさらに含む、請求項44記載の方法。

46. 対象が特定の疾患または障害を有する人々であり、属性が疾患または障害に対する寄与因子の可能性があるものである、請求項2記載の方法。

47. 対象が多数の異なる疾患または障害を有する人々であって属性が該疾患または障害に対する寄与因子の可能性があるものである、請求項2記載の方法。

48. 対象が疾患または障害に対する寄与因子の可能性があるものを含み属性がそれらの因子を持つ人々または持たない人々であって、該疾患または障害に対する実質的に等価なリスクを持つ人々の群を関連づける、請求項2記載の方法。

49. 対象が時間刻みであって属性がシステムの故障前の時間刻みでのシステム内の要素の状態を含み、システムの故障を潜在的に引き起こしうる要素の状態を関連づける、請求項2記載の方法。

50. r_i がすべての反復に関して同一である、請求項1記載の一致検出法。

51. 第１に、システム状態が選択された時間量にわたる状態変数の値によって提示されるシステム状態間の移行のデータベースを作成する段階、および各状態から状態への移行の集合がM個の対象の一つに対応し、このため各状態変数がある属性に対応するようなデータセットとして全体的または部分的にデータベースを提示する段階をさらに含む、請求項2記載の方法。

52. 第１に、選択された時間量にわたる状態および作用のデータベースを作成す

る段階、および各状態／作用／状態の3つ組がM個の対象の一つに対応し、このため各状態変数または作用のタイプがある属性に対応するようなデータセットとして全体的または部分的にデータベースを提示する段階をさらに含む、請求項2記載の方法。

53. 対象対属性という行列の形式で表現された多数の属性を有する対象のデータセットとともに用いるための一致検出法であって、下記の段階を含む方法：

・各反復で行列のサンプリングされた部分集合が各対象に関して属性の同じ部分集合を有する、あらかじめ決定された数の反復にわたる行列の部分集合のサンプリング、

・一致が行列のサンプリングされた部分集合における1つまたは複数の対象における複数の属性値の同時出現であって、複数の属性値が各出現に関して同一であって、サンプリングされた各部分集合における一致の検出および数の記録が他の部分集合における一致のサンプリング、検出および数の記録の前、同時または後に行われる、行列のサンプリングされた各部分集合における一致の検出および数の記録、

・サンプリング、検出および記録の前、同時または後に行われる、関心対象の各一致に関する期待数の決定、

・関心対象の各一致に関する一致の観測数と一致の期待数との比較、およびこの比較による、一致に関する複数の属性の相関の程度の決定、ならびに

・k項数の相関属性が、それに関する相関の程度がそれぞれにあらかじめ決定され

た閾値を上回る複数の属性である、k項数の相関属性の集合の報告。

54. 数値的相関値がk項数の相関属性の集合とともに報告される、請求項1記載の方法。

【発明の詳細な説明】

一致検出の方法、製品および装置

技術分野

本発明は、多数の変数間の一致を検出するための方法、装置およびシステムに関する。さらに本発明は、種々の分野への一致検出法の適用、およびこのような適用に由来する製品に関する。

背景技術

k項数（k-tuple）の相関属性（correlated attributes）

対（pair）またはk項数（k-tuple）の変数間の相関の発見には、科学、医学、工業および商業の多くの領域で用途がある。例えば、医師および公衆衛生専門家にとって、いかなる生活様式、食事および環境要因が互いに、ならびに患者の病歴データベースにおける特定の疾患と相関するかを知ることには大きな関心がある。株または商品の取引業者にとって、価格が時とともに共変する一組の金融証書を見いだすことは収益につながる可能性がある。スーパーマーケットチェーンまたは通信販売会社の販売損は、製品Aを購入する消費者に製品BおよびCも購入する傾向があるかを知ることに興味を抱くと思われ、これは販売記録データベースに見いだしうる。数理分子生物学および薬物探索の研究者は、整列化されたRNAまたは蛋白質配列のセットにおける離れた配列要素間の相関から3D分子構造の諸性質を推測したいと考えるであろう。

多くの多様な用途を包含し、本明細書に記載される原理の理解を促す一般的な問題を定式化したものの1つが、行が「対象（object）」（個々の患者、株価、消費者または蛋白質配列など）に対応し、列が特徴または属性または変数（生活様式の諸因子、株、販売品目またはアミノ酸残基の位置など）に対応する、離散的特徴を持つ行列である。

任意の2つ、またはさらには3つもしくは4つの特定の変数の間の相関の種類、度合いおよび統計的有意性の程度を決定するための数学的方法は一般化しており、よく知られている。これらの方法には、連続変数に関する線形および非線形回帰ならびに離散変数に関する分割表分析が含まれる。しかし、はるかに大規模な変数の集合に関する相関を評価しようとすると―または同時確率もしくは条件付

き

確率を評価しようとしただけでも一大きな困難が生じる。この扱いにくさには結合属性値確率密度項（joint attribute-value probability density term）が多すぎるという1つの主な原因があり、これ自体が以下の2つの重大な問題を明らかに示す：（1）データベース全体にわたってすべての項に関して度数を計算および保存するには非常に多くの計算およびメモリを要する、（2）これらの度数に基づいて信頼性のある確率推定値を提供するにはデータベースのレコード数が通常は不十分である。

少し詳しく考察してみよう。M個のレコード（対象）、N個の変数（属性、フィールド）に関して、各変数が|A|通りの値をとりうる同一の集合をなすと仮定すると、

$\binom{N}{k}=N!/(N-k)!k!$ 通りのk項数の列が得られる。$k=1, 2, \cdots, N_A$のそれぞれに関してk項数の数を加えると、すべてのサイズに関するこうした項数は2^N-1個となる。この指数関数的な複雑さが、高次確率推定および相関検出の方法の主な障害となっていた。

この複雑さに対する1つの自然な考え方は、列変数（column varibale）の集合の幕集合（power set）である。この幕集合は演算⊂の下で、結節点（node）がこの列変数の集合の部分集合となるグラフに対応する「塔（tower）」である数学的格子（mathematical lattice）を形成する（集合にN個の要素があれば、幕集合は2^N個の要素を持つことに注意）。この観点からは、部分集合σ_1およびσ_2を表す2つの結節点が連結されるのは$\sigma_1 \subset \sigma_2$または$\sigma_2 \subset \sigma_1$の場合であり、しかもこれらの場合のみである。$\sigma_1 \subset \sigma_2$の場合、本発明者らは$\sigma_2$の結節点は$\sigma_1$のものより上にあると表現する。これにより「高次」という用語に対して、塔のより上方にあるという自然な意味が生じる。本発明者らは底すなわち零集合結節点（nullset node）を0番目の層と呼ぶ。単一列項は1番目の層をなし、以下も同様である。

塔によるたとえをさらに続け、本発明者らは、この建造物の各「フロア」は

$\binom{N}{k}$ 個の「スイート」を含み、各スイートは$|A|^k$個の「部屋」を含むと表現する。言い換えると、格子のk階は $\binom{N}{k}$ 通りの異なるk項数の列変数に対応し、各

k項数は（|A|by|A|…by|A|）の分割表と関連し、その各セルにはそれらの特

定のk列の間の相関に関する古典的分割表検定に用いるための特定の結合記号（joint symbol）（a_{i1}，a_{i2}，…，a_{ik}）の計数度数が収められている必要がある（図1参照）。

任意のk∈｛1，2，…，N｝、任意の特定のk項数の列（c_{j1}，c_{j2}，…，c_{jk}）に関して、とりうる結合値（joint value）は|A|個ある。任意のk∈｛1，2，…，N｝、任意の特定のk項数の列（c_{j1}，c_{j2}，…，c_{jk}）に関して、このデータセットを用いるカルバック発散またはその他の相関関数の推定は少なくともΩ(Mk)またはΩ($|A|^k$)量の計算であり、M、kおよび|A|の相対的サイズに依存する。

データベースの包括的確率論モデルは、 $\Sigma_{k=1}^{N}\binom{N}{k}|A|^k$ 個の項に関

する確率推定値を特定しうる必要がある。このことは、例えば数理分子生物学の領域では、各配列が7つのアミノ酸残基を有する小さなヘプタペプチド配列ファミリーに関して特定される項は1,801,088,540個あることを意味する。4つの塩基記号という少数のRNA用アルファベットで書かれた長さ15ヌクレオチドの非現実的に小さなRNAでも30,517,578,124個もの項がある。

明らかにこうしたモデルは取り扱えないほど巨大化しうる。モデル化／学習手順によって検索しなければならない可能なモデルの領域についてはどうであろうか。その状態が顕在変数（observable）に共同的に影響を及ぼす潜在変数（latent variable）を仮定することによって顕在変数の集合間の相関を説明しようとする潜在変数モデルについて考察する。各モデルはk項数の変数の集合を特定しなければならず、このような集合はexp(2，2^N)（すなわち$2\sim2^N$乗）個存在するため、最悪の場合には検索領域にexp(2，2^N)個の可能なモデルが存在することになる。

高次確率の程度を決定するための種々の方法では、求める高次特徴の幅k（図3

参照）、位置（図2）、数または相関の度合い、および考慮するモデルの種類に対してあらかじめ厳密な制限を加えることにより、組み合わせの爆発的増加を回避しようとする（図4参照）。

確率推定の3つの目標

既存の方法および本発明の詳細を説明する前に、それぞれが多くの研究および今日の常法に対応する、大規模なデータセットにおける確率推定に関して考えうる3つの異なる目標を概説しておくことが有用である。

１．完全に特定された完全に高次の同時確率分布の推定：すべてのk項数の属性およびとりうる値に関して

$q(a_{i1}@c_{i1}, a_{i2}@c_{i2}, \cdots, a_{ik}@c_{ik})$

を特定する確率密度qを推定する。

２．特定の属性および特定の変数に関する特定の仮説に関する仮説の検定：例えば、データは列$c_{i1}, c_{i2}, \cdots, c_{ik}$が独立であるとの仮説と一致するか？。

３．特徴検出または「データマイニング（data mining）」：最も疑われる一致、例えば、低次限界から予測されるであろうものよりも確率が高い同時属性出現を検出する。これに関連して、最も相関性の高いk項数の列を見いだす。

本発明と最も関連の深いものは特徴検出およびデータマイニングの用途である。しかし、データベースの完全な高次同時確率分布を推定するための最も好首尾な方法の中には、$k \geq 2$の変数の集合間で高い相関を示すような高次項の正確な特定、および最大エントロピーの仮定を要するものがあり、したがってそのような用途も本発明の目的である。

関連する研究

高次確率の推定、相関の検出、および高次データベース関連のモデル化のために、種々の数学的および計算的方法が提唱され、用いられている。従来のすべてのこのような方法は、非常にコストのかかるすべての可能なk項数の変数にわたる全体的、時には網羅的な検索を用いるか、または特別に固定した少数のkのk項数のみにそれらの検索を限定することにより複雑性を完全に回避する（しばしば$k=2$であり、このため考慮されるのは対相関のみである）かのいずれかである。

関連する研究の代表的な例を以下にいくつか列挙する。

属性間の独立性の仮定　高次相関の複雑さを回避する最も簡単な方法は、単にそれらが存在しないとみなすことである。本方法の用途のいくつかの分野において歴史的に主流である多くのアルゴリズムおよびコンピュータプログラムは、単にすべての変数、すべての属性が独立であるというデータのモデルを作成および使用している。例えば、数理分子生物学におけるDNAおよび蛋白質配列のモデル化は、異なる塩基またはアミノ酸残基の位置は独立であるとの誤った仮定に立っており、しばしば共通配列およびプロファイルを用いてなされる。このようなモデ

ルに依拠すると、モデル化しようとするDNAまたは蛋白質に関する極めて重要な機能的および構造的洞察を覆い隠すおそれがある。

kに関する事前の制限　データベースのギブスのモデルに関する1つの提唱はギブスポテンシャルの使用に基づき、これらの特殊な項を計算するためのハッシュ法（hashing method）を提唱している。各k次ポテンシャルには、k次の同時確率密度に加えて若干数の低次（典型的にはk-1次）密度の推定が必要である。ポテンシャル計算の主な成分であるミラーのパターン収集サブルーチンの漸近的時間複雑性は、本発明者らの用語に解釈すると以下のようになる：

$$M \cdot \sum_{k=1}^{K} \binom{N_A}{k} 2^k \approx O(MN^K)$$

ここで$K=k_{max}$は、それに関して検索しようとし、それによってデータベースのオブジェクトが示されると思われる最も高次の特徴である。この指数関数的な爆発的増加は、数百もの属性を有するデータベースにおける、4または5を大きく上回る任意のk次の高次特徴（HOF）に関する検索を妨げる。

さまざまな適用領域における多くの方法では、単にkをk＝2と限定している。例えば、対残基間相関法（pairwise inter-residue correlation method）は、蛋白質の構造および機能の予測に有用な可能性があって一次配列分類機およびフォールド分類機よりも感度の高い分類機に組み込みうる二次特徴を見いだす。k項（k-ary）相互作用が重要である限り、およびこのような相互作用が相同配列

の集合に痕跡を残す限りにおいて、対による方法には欠陥がある。2項相関の集合からk項相関を推定しようと試みることはできるが［9］（本質的には「CorrelatesWith」二値関係の推移的閉鎖の計算による）、この帰納的方法はトラブルを招く恐れがある:すなわち、変数x、y、zの間に高い対相関が認められてもそれは一般にはx、y、zの3変数の3項相関（カルバック発散によって計測されるようなもの）の高さを意味せず、その逆も必ずしも成り立たない。多剤相互作用の検討などのその他の応用領域でも、対相関検出法では重大な高次関連が見逃される恐れがあることは同様にいえる。

無作為変数の最も相関する対を同定するためのパツリ（Paturi）らの方法　N個のランダムな二値変数X_1，X_2,…，X_Nの大きな集合の中から最も相関する変数の対

X_i、X_jを見いだす問題に関して1つの方法が報告されている。この方法は、最も相関するk項数の無作為二値変数の発見へと容易に拡張しうるが、計算上の複雑性は著しく増大し、しかもこれは先験的に固定されたk≧2のみに関する。ここでは、M個の標本$\{X^m_1, X^m_2, \cdots, X^m_N\}_{m=1,2,\ldots,M}$　の何らかの集合にわたる相関$(X_i, X_j)=P\{X_i=X_j\}$という相関の定義を用いている（ここで$P[X_i=X_j]$は、変数X_iが変数X_jと同じ値または状態を有する確率を意味する）。彼らの方法では、2つまたはそれ以上のほぼ等しく相関する対（またはk項数）の変数を分離しようとする際に、時間的複雑性および標本の複雑性の両面でかなりの計算的複雑性を招く。

パツリ法の2つの変数はNにおいてそれぞれ漸近的に二次的および準二次的（sub-quadratic）であり、手順を迅速化するにはより多くの標本抽出を必要とする。この方法を最大のk項相関の探索に拡張する場合には（ここでは相関は$P[X_{i1}=X_{i2}=\cdots X_{ik}]$と定義される）、時間的複雑性はほぼ$O(k^2N^k\log^3N)$に増大する。極めて大きなデータセットにおいて5または6を大きく上回る幅kの高相関属性の小集団を検索することは、ここでも同じく除外される。

隠れマルコフモデル　隠れマルコフモデル（HMM）は、自動会話認識ならびに蛋白質、DNAおよびRNA配列のモデル化の両面において近年広くしかもますます好

首尾に用いられている。

いくつかのグループがHMMによる蛋白質配列ファミリーおよび連続会話データのモデル化の大きな成功を報告しているが、あらかじめ選択された高次特徴のHMMへの「直結（hardwiring）」による学習時間およびモデルの頑健性（robustness）の点では改良の余地は大きい（これは種々の分野においてHMM様反回神経回路網に関して検討されている）。

そもそもHMMが局所的配列相関を用いる蛋白質配列または記録された発話の整列化に極めて優れるという理由のいくつかは、同時に、すでに部分的または完全に整列化されたデータにおける重要な配列-距離相関を見いだす上でのこのような方法の有用性の低さにもつながる。このジレンマに起因する現象は「拡散（diffusion）」と呼ばれる。

一次HMMでは定義により配列の列（sequence column）の独立性を仮定しており、隠れ状態配列が与えられる。比較的長い範囲の相互作用を捕捉するためには原則的には複数の選択的状態配列を用いうるが、k項数の相関列の数に伴い、これらの数は指数関数的に増加する。

結合規則を発見するためのアグローワル（Agrawal）らの方法　この方法は、データベースから知識に基づく規則を自動抽出するという、おそらく最も純粋なデータマイニングの文脈で開発されたものである。この方法では、M個のトランザクション（対象、行）およびN個の項目（属性、列）のデータベースを考慮し、a⇒bの形式の規則を抽出しようとする。このため、「aを含むトランザクションがbを含む傾向がある」というような属性a、bの対、すなわちp(b|a)が高値をとるような対を探索する。「CDプレーヤーを購入する人々はCDを購入する傾向がある」とは、このような方法における商業的な利益の可能性を示唆する一例である（より一般的には、$p(b_1, b_2, \cdots, b_k | a_1, a_2, \cdots, a_j)$が高値である属性の集合に関して検索しうる）。

規則a⇒bは以下を有するとされている：

１．aを含むトランザクションのc%がbも含む場合（したがって、大まかには$p(a,b)/p(a) \geqq (c/100)$であれば）、信頼度（confidence）c；

２．aおよびbを含むトランザクションがs％であれば（したがって、大まかにはp(a, b)≧s／100であれば）、サポート（support）s。

この方法の背景にある目標は本発明の目的とは異なる。しかし、アグローワル（Agrawal）の方法による対称性規則の発見に目的を絞り（p(a, b)／aおよびp(a, b)／bがいずれも高値を示す属性の対に関して検索がなされるように）、サポートに対する強調を抑制すれば（稀に起こるものであっても疑わしい一致が検索されるように）、異なる目的もより密接に結びつく。

アグローワルの方法では、‖S‖をこの手順において特定の段階の処理に到達する指数関数的に大きい数のk項数（1≦k≦Nの任意のサイズに関する）αの属性に関するSupport(α)のすべての値の合計とすると、複雑さが$O(\|S\|-MN)$倍になることが示されている。したがって、この方法は最悪の場合には$O(2^N)$の量となる。それぞれの分野で現実的なデータセットと考えられるものに対して一連の経験的検定が行われる。この手順の実行時間はトランザクションの数Mに伴って線形的に増加するに過ぎないが、項または属性の数は$N_A=1000$に一定に保たれ、構成され

たそれらのデータセットは幅k＞10の相関するk項数をおそらくは含まない。k-1次の小集団（clique）からのk次の小集団の漸増的構築に基づくそれらのアルゴリズムの解析から、この方法では幅広いHOF（大きいk）を見いだすために、同等の統計的有意性を有するより狭いHOF（より小さなk）よりもはるかに多くの計算を行うことが明らかとなっている。

スティーグ、ロビンソン、ディアフィールド、ラッパ（Steeg、Robinson、Deerfield、Lappa）－1993　整列化された蛋白質配列の集合におけるk項数の相関残基（位置）を見いだすことを目的として、大まかで経験的な方法がいくつか提唱されている。提示された方法の1つは、本明細書に記載される表現および一致検出の諸段階の初歩的なものの一態様を用いている。

属性間の相関を見いだすための代替的な方法および装置、ならびにそれらの相関に関する応用が必要である。

発明の開示

第1の局面において、本発明は、多数の属性を有する対象のデータセットとともに用いるための一致検出法を提供する。基本的方法は以下の段階を含む：

・対象がある属性を有する場合にその属性が対象において出現すると呼ばれる、N_A個の変数（「属性」）を項とするM個の対象の集合の表示、

・あらかじめ決定された数の反復（iteration）における各反復に関する、M個の対象からのr_i個の部分集合のサンプリング、

・一致がサンプリングされた部分集合におけるr_i個の対象のうち同じh_i個における$1 \leqq k \leqq N_A$個の属性の同時出現であって、$0 \leqq h_i \leqq r_i$である、サンプリングされた対象の部分集合のそれぞれにおけるk個の属性の集合間での一致の検出および記録、

・決定がサンプリングおよび収集の前、同時またはサンプリングおよび収集の後に行われる、上記のk個の属性の任意の集合ならびにあらかじめ決定された数のサンプリングおよび一致計数の反復に関する一致の期待数の決定、

・k個の属性の任意の集合ならびにサンプリングおよび一致計数の反復回数に関する一致の観測値と期待数との比較、ならびにこの比較によるk個の属性の集合に関する相関（または結合または依存）の程度の決定、ならびに

・k項数の相関属性が、選択された相関の程度に関してあらかじめ決定された閾値を上回る値を持つことがこの過程によって決定されたN_A個の属性のkの集合である、k項数の相関属性の集合の報告。

第2の局面において、本発明は、多数の属性を有する対象のデータセットとともに用いるための一致検出法であって、下記の段階を含む方法を提供する：

・各反復でデータセットのサンプリングされた部分集合が各対象に関して属性の同じ部分集合を有する、あらかじめ決定された数の反復にわたるデータセットの部分集合のサンプリング、

・一致がデータセットのサンプリングされた部分集合における1つまたは複数の対象における複数の属性値の同時出現であって、複数の属性値が各出現に関して同一であって、データセットのサンプリングされた各部分集合における一致の検出および数の記録が他の部分集合における一致のサンプリング、検出および数の

記録の前、同時または後に行われる、データセットのサンプリングされた各部分集合における一致の検出および数の記録、

・サンプリング、検出および記録の前、同時または後に行われる、関心対象の各一致に関する期待数の決定、

・関心対象の各一致に関する一致の観測数と一致の期待数との比較、およびこの比較による、一致に関する複数の属性の相関の程度の決定、ならびに

・k項数の相関属性が、それに関する相関の程度がそれぞれにあらかじめ決定された閾値を上回る複数の属性である、k項数の相関属性の集合の報告。

いずれの局面においても、観測数と期待数との比較は尾部確率（tail probability）に関するチャーノフ境界（Chernoff bound）を用いて計算することができ、数はサンプリングされた部分集合のすべてにわたって各一致の数の連続的な合計を保存することによって記録しうる。

第3の局面において、本発明は、多数の属性を有する対象のデータセットの視覚的表示のための方法であって、下記の段階を含む方法を提供する：

・各反復でデータセットのサンプリングされた部分集合が必ずしも同一の対象である必要はないものの同じ数の対象を有し、各対象に関して属性の同じ部分集合を有する、あらかじめ決定された数の反復にわたるデータセットの部分集合のサンプリング、

・一致がデータセットのサンプリングされた部分集合における1つまたは複数の対象における複数の属性値の同時出現であって、複数の属性値が各出現に関して同一であって、データセットのサンプリングされた各部分集合における一致の検出および数の記録が他の部分集合における一致のサンプリング、検出および数の記録の前、同時または後に行われる、データセットのサンプリングされた各部分集合における一致の検出および数の記録、

・サンプリング、検出および記録の前、同時または後に行われる、関心対象の各一致に関する期待数の決定、

・関心対象の各一致に関する一致の観測数と一致の期待数との比較、およびこの比較による、一致に関する複数の属性の相関の程度の決定、ならびに

・k項数の相関属性が、それに関する相関の程度がそれぞれにあらかじめ決定された閾値を上回る複数の属性である、k項数の相関属性の集合のグラフィカルインターフェースを通じてのユーザーに対する報告。

第4の局面において、本発明は、多数の属性を有する対象のデータセットの高次相互作用を捕捉してデータモデル化ユニット（data modelling unit）に報告するための、データモデル化ユニットとともに用いるための予備処理の方法であって、下記の段階を含む方法を提供する：

・各反復でデータセットのサンプリングされた部分集合が各対象に関して属性の同じ部分集合を有する、あらかじめ決定された数の反復にわたるデータセットの部分集合のサンプリング、

・一致がデータセットのサンプリングされた部分集合における1つまたは複数の対象における複数の属性値の同時出現であって、複数の属性値が各出現に関して同一であって、データセットのサンプリングされた各部分集合における一致の検出および数の記録が他の部分集合における一致のサンプリング、検出および数の記録の前、同時または後に行われる、データセットのサンプリングされた各部分集合における一致の検出および数の記録、

・サンプリング、検出および記録の前、同時または後に行われる、関心対象の各一致に関する期待数の決定、

・関心対象の各一致に関する一致の観測数と一致の期待数との比較、およびこの比較による、一致に関する複数の属性の相関の程度の決定、ならびに

・k項数の相関属性が、それに関する相関の程度がそれぞれにあらかじめ決定された閾値を上回る複数の属性である、k項数の相関属性の集合の、データモデル化ユニットへの報告。

第5の局面において、本発明は、多数の属性を有する対象のデータセットとともに用いるための相関消去の方法であって、下記の段階を含む方法を提供する：

・各反復でデータセットのサンプリングされた部分集合が各対象に関して属性の同じ部分集合を有する、あらかじめ決定された数の反復に関するデータセットの部分集合のサンプリング、

・一致がデータセットのサンプリングされた部分集合における1つまたは複数の対象における複数の属性値の同時出現であって、複数の属性値が各出現に関して同一であって、データセットのサンプリングされた各部分集合における一致の検出および数の記録が他の部分集合における一致のサンプリング、検出および数の記録の前、同時または後に行われる、データセットのサンプリングされた各部分集合における一致の検出および数の記録、

・サンプリング、検出および記録の前、同時または後に行われる、関心対象の各一致に関する期待数の決定、

・関心対象の各一致に関する一致の観測数と一致の期待数との比較、およびこの比較による、一致に関する複数の属性の相関の程度の決定、ならびに

・k項数の相関属性が、それに関する相関の程度がそれぞれにあらかじめ決定された閾値を上回る複数の属性である、k項数の相関属性の集合の消去。

いずれの局面においても、対象が各トランザクションが1つまたは複数の購入された製品を含む販売トランザクションであって、属性が特定の製品または特定の種類の製品の販売の事例であってもよい。対象が時間刻みであって、属性がシステムにおける要素の状態であってもよい。対象が時間刻みであって、属性が金融証書または商品の価格または価格の変動であってもよい。

いずれの局面においても、本方法の段階は以下の疑似コードによって表現されうる：

```
0. begin
1. read(MATRIX);
2. read(R, T);
3. compute_first_order_marginals(MATRIX);
4. csets:={};
5. for iter=1 to T do
6. sampled_rows:=rsample(R, MATRIX);
7. attributes:=get_attributes(sampled_rows);
8. all_coincidences:=find_all_coincidences(attributes);
```

```
9. for coincidence in all_coincidences do
10. if cset_already_exists(coincidence, csets)
11. then update_cset(coincidence, csets);
12. else add_new_cset(coincidence, csets);
13. endif
14. endfor
15. endfor
16. for cset in csets do
17. expected:=compute_expected_match_count(cset);
18. observed:=get_observed_match_count(cset);
19. stats:=update_stats(cset, hypoth_test(expected, observed));
20. endfor
21. print_final_stats(csets, stats);
22. end
```

第6の局面において、本発明は、各対象が複数の属性を有する、対象のデータセットとともに用いるための一致検出システムであって、下記の手段を含むシステムを提供する：

・各反復でデータセットのサンプリングされた部分集合が各対象に関して属性の同じ部分集合を有する、あらかじめ決定された数の反復にわたるデータセットの部分集合のサンプリングのための手段、

・一致がデータセットのサンプリングされた部分集合における1つまたは複数の対象における複数の属性値の同時出現であって、複数の属性値が各出現に関して同一であって、データセットのサンプリングされた各部分集合における一致の検出および数の記録が他の部分集合における一致のサンプリング、検出および数の記録の前、同時または後に行われる、データセットのサンプリングされた各部分集合における一致の検出および数の記録のための手段、

・サンプリング、検出および記録の前、同時または後に行われる、関心対象の各一致に関する期待数の決定のための手段、

・関心対象の各一致に関する一致の観測数と一致の期待数との比較、およびこの比較による、一致に関する複数の属性の相関の程度の決定のための手段、ならびに

・k項数の相関属性が、それに関する相関の程度がそれぞれにあらかじめ決定された閾値を上回る複数の属性である、k項数の相関属性の集合の報告のための手段。

第6の局面のシステムにおいて、データセットの部分集合のサンプリングのための手段は、データセットを分割してサンプリング用の部分集合にするための手段を含みうる。一致の検出および数の記録のための手段は、それぞれのプロセッシングノード（processing node）が一致の検出および各サブカウントの記録を行うプロセッシングノードのアレイを含むことができ、関心対象の各一致に関して前記一致の観測数と前記一致の期待数とを比較するための手段は、前記サブカウントをマージ（merge）して前記観測数を提供するための手段を含みうる。前記プロセツシングノードの少なくとも1つは一致の検出およびそれぞれのサブサブカウントの記録を行うそれぞれのプロセッシングノードのサブアレイを含むことができ、マージのための前記手段は前記サブサブカウントをマージして前記サブカウントおよび／または前記観測数を提供する。各プロセッシングノードは、データセットの受け取られた部分集合を保存するための入力バッファーおよびサブカウントまたはサブサブカウントを保存するための出力バッファーを含むメモリ、ならびにメモリとの間でデータをやり取りするメモリバスを含みうる。

第7の局面において、本発明は、コンピュータおよび多数の属性を有する対象のデータセットとともに用いるための一致検出プログラム媒体であって、

そのコンピュータと互換性がある保存媒体上に保存されたコンピュータプログラムであって、

・各反復でデータセットのサンプリングされた部分集合が各対象に関して属性の同じ部分集合を有する、あらかじめ決定された数の反復にわたるデータセットの部分集合のサンプリング、

・一致がデータセットのサンプリングされた部分集合における1つまたは複数

の対象における複数の属性値の同時出現であって、複数の属性値が各出現に関して同一であって、データセットのサンプリングされた各部分集合における一致の検出および数の記録が他の部分集合における一致のサンプリング、検出および数の記録の前、同時または後に行われる、データセットのサンプリングされた各部分集合における一致の検出および数の記録、

・サンプリング、検出および記録の前、同時または後に行われる、関心対象の各一致に関する期待数の決定、

・関心対象の各一致に関する一致の観測数と一致の期待数との比較、およびこの比較による、一致に関する複数の属性の相関の程度の決定、ならびに

・k項数の相関属性が、それに関する相関の程度がそれぞれにあらかじめ決定された閾値を上回る複数の属性である、k項数の相関属性の報告

のためにコンピュータを指向するための指示を含むコンピュータプログラムを含む媒体を提供する。

第8の局面において、本発明は、多数の属性を有する対象のデータセットとともに用いるための一致検出システムであって、

コンピュータ、ならびに

そのコンピュータと互換性のある媒体上のコンピュータプログラムであって、

・各反復でデータセットのサンプリングされた部分集合が各対象に関して属性の同じ部分集合を有する、あらかじめ決定された数の反復にわたるデータセットの部分集合のサンプリング、

・一致がデータセットのサンプリングされた部分集合における1つまたは複数の対象における複数の属性値の同時出現であって、複数の属性値が各出現に関して同一であって、データセットのサンプリングされた各部分集合における一致の検

出および数の記録が他の部分集合における一致のサンプリング、検出および数の記録の前、同時または後に行われる、データセットのサンプリングされた各部分集合における一致の検出および数の記録、

・サンプリング、検出および記録の前、同時または後に行われる、関心対象の

各一致に関する期待数の決定、

・関心対象の各一致に関する一致の観測数と一致の期待数との比較、およびこの比較による、一致に関する複数の属性の相関の程度の決定、ならびに

・k項数の相関属性が、それに関する相関の程度がそれぞれにあらかじめ決定された閾値を上回る複数の属性である、k項数の相関属性の集合の報告

のためにコンピュータを指向するためのコンピュータプログラム

を含むシステムを提供する。

いずれの局面においても、本発明の方法は、データセットのサンプリングの前に、データセットがその行列のサンプリングによってサンプリングされる、対象と属性からなる行列の形で対象および属性を提示する段階をさらに含みうる。

第9の面において、本発明は、

・各反復でデータセットのサンプリングされた部分集合が必ずしも同一の対象である必要はないものの同じ数の対象を有し、各対象に関して属性の同じ部分集合を有する、あらかじめ決定された数の反復に関する対象対属性を表現するデータセットの部分集合のサンプリング、

・一致がデータセットのサンプリングされた部分集合における1つまたは複数の対象における複数の属性値の同時出現であって、複数の属性値が各出現に関して同一であって、データセットのサンプリングされた各部分集合における一致の検出および数の記録が他の部分集合における一致のサンプリング、検出および数の記録の前、同時または後に行われる、データセットのサンプリングされた各部分集合における一致の検出および数の記録、

・サンプリング、検出および記録の前、同時または後に行われる、関心対象の各一致に関する期待数の決定、

・関心対象の各一致に関する一致の観測数と一致の期待数との比較、およびこの比較による、一致に関する複数の属性の相関の程度の決定、ならびに

・k項数の相関属性が、それに関する相関の程度がそれぞれにあらかじめ決定された閾値を上回る複数の属性である、k項数の相関属性の集合の報告

によって選択される属性の集合を有する製品を提供する。

第10の局面において、本発明は、

・各反復でデータセットのサンプリングされた部分集合が各対象に関して属性の同じ部分集合を有する、あらかじめ決定された数の反復に関する対象対属性を表現するデータセットの部分集合のサンプリング、

・一致がデータセットのサンプリングされた部分集合における1つまたは複数の対象における複数の属性値の同時出現であって、複数の属性値が各出現に関して同一であって、データセットのサンプリングされた各部分集合における一致の検出および数の記録が他の部分集合における一致のサンプリング、検出および数の記録の前、同時または後に行われる、データセットのサンプリングされた各部分集合における一致の検出および数の記録、

・サンプリング、検出および記録の前、同時または後に行われる、関心対象の各一致に関する期待数の決定、

・関心対象の各一致に関する一致の観測数と一致の期待数との比較、およびこの比較による、一致に関する複数の属性の相関の程度の決定、ならびに

・k項数の相関属性が、それに関する相関の程度がそれぞれにあらかじめ決定された閾値を上回る複数の属性である、k項数の相関属性の集合の報告

によって生成される一組の規則を適用することによって規定される製品を提供する。

いずれの局面においても、本発明の方法は、報告された相関属性によって規定される規則を適用する段階をさらに含みうる。

第11の局面において、本発明は、残基A18/Q31/H33の空間座標を含むHIVエンベロープ蛋白質のV3ループの構造モチーフを含むペプチドまたは疑似ペプチド（peptidomimetic）を提供する。

第12の局面において、本発明は、請求項2記載の方法を用いて同定される構造モチーフを有する蛋白質と相互作用するリガンドを含む薬学的組成物、およびそのための薬学的に許容される担体または賦形剤を提供する。リガンドは適した実体

を有し、その成分が対応するモチーフの残基または部分と相互作用するような位

置に互いにある化学的成分を含みうる。リガンドはモチーフとの相互作用により、モチーフを含む蛋白質の領域の機能を妨げうる。

第13の局面において、本発明は、本発明のこれまでの局面の方法を用いて同定される構造モチーフを有する蛋白質と相互作用するリガンド、およびそのリガンドと結合した検出可能な標識を含む診断薬を提供する。

第14の局面において、本発明は、残基A18/Q31/H33の空間座標を有するV3ループの構造モチーフを含むエンベロープ蛋白質であって、そのモチーフと相互作用する官能基を少なくとも1つ含むリガンドを含有する、ヒト免疫不全ウイルス（HIV）のエンベロープ蛋白質と相互作用する薬学的組成物、ならびにそのための薬学的に許容される担体または賦形剤を提供する。リガンドは、残基18との結合能を有していて前記リガンド中の残基18との結合のための有効部分に存在する少なくとも1つの官能基、残基31との結合能を有していて前記リガンド中の残基31との結合のための有効部分に存在する少なくとも1つの官能基、および残基33との結合能を有していて前記リガンド中の残基33との結合のための有効部分に存在する少なくとも1つの官能基を含みうる。

第15の局面において、本発明は、ヒト免疫不全ウイルス（HIV）のエンベロープ蛋白質の構造モチーフと相互作用するリガンドを設計する方法であって、HIVエンベロープ蛋白質のV3ループ内の残基A18、Q31およびH33の空間座標を有するテンプレートの提供、ならびに空間的拘束を有する有効なアルゴリズムを用いての化学的リガンドの計算による展開（evolve）であって該展開されたリガンドがモチーフと結合する有効な官能基を少なくとも1つ含むような展開、の段階を含む方法を提供する。リガンドは、残基18との結合能を有していて前記リガンド中の残基18との結合のための有効部分に存在する少なくとも1つの官能基、残基31との結合能を有していて前記リガンド中の残基31との結合のための有効部分に存在する少なくとも1つの官能基、および残基33との結合能を有していて前記リガンド中の残基33との結合のための有効部分に存在する少なくとも1つの官能基を含みうる。

第16の局面において、本発明は、ヒト免疫不全ウイルス（HIV）のエンベロープ蛋白質の構造モチーフと結合するリガンドを同定する方法であって、HIVエン

ベロ

ープ蛋白質のV3ループ内の残基A18、Q31およびH33の空間座標を有するテンプレートの提供、分子の構造および配向性を含むデータベースの提供、ならびにその成分がモチーフと相互作用するように互いに対して位置する有効成分を前記分子が含むかどうかを決定するための前記分子のスクリーニング、の段階を含む。分子の第1の成分が残基18と相互作用し、分子の第2の成分が残基31と相互作用し、分子の第3の成分が残基33と相互作用することが可能である。

第17の局面において、本発明は、本明細書に記載される共変するk項数を具現化した抗原およびワクチンを提供しうる。

第18の局面において、本発明は、

・各反復でデータセットのサンプリングされた部分集合が必ずしも同一の対象である必要はないものの同じ数の対象を有し、各対象に関して属性の同じ部分集合を有する、あらかじめ決定された数の反復に関する対象対属性を表現するデータセットの部分集合のサンプリング、

・一致がデータセットのサンプリングされた部分集合における1つまたは複数の対象における複数の属性値の同時出現であって、複数の属性値が各出現に関して同一であって、データセットのサンプリングされた各部分集合における一致の検出および数の記録が他の部分集合における一致のサンプリング、検出および数の記録の前、同時または後に行われる、データセットのサンプリングされた各部分集合における一致の検出および数の記録、

・サンプリング、検出および記録の前、同時または後に行われる、関心対象の各一致に関する期待数の決定、

・関心対象の各一致に関する一致の観測数と一致の期待数との比較、およびこの比較による、一致に関する複数の属性の相関の程度の決定、ならびに

・k項数の相関属性が、それに関する相関の程度があらかじめ決定された閾値を上回る複数の属性である、k項数の相関属性の集合の報告

によって選択される一組の属性との相互作用によって規定される製品を提供する。

いずれの局面においても、対象は化合物であってもよく、属性が特定の化学成分を含んでいてもよい。対象がペプチドまたは蛋白質であり、属性がモチーフの特定の構造または下部構造のパターンを含んでもよい。対象が化合物、分子構造、ヌクレオチド配列およびアミノ酸配列からなる群より選択され、属性が選択された対象の特徴であってもよい。対象が時間刻みであって属性が遺伝子または遺伝子産物の生物的パラメーターであってもよい。対象が電子的に保存される、および／または電子的に索引が付けられた（indexed）文書であり、属性が題目であってもよい。対象が消費者であって属性がそれらの消費者によって購入された、または購入されなかった製品を含んでもよい。属性が、消費者に対して郵送されたこと、またはされなかったことをさらに含んでもよい。対象が製品を含むものであって属性がそれらの製品を購入した、または購入しなかった消費者を含んでいてもよい。属性が消費者の人口統計変数をさらに含んでもよい。対象が特定の疾患または障害を有する人々であり、属性がその疾患または障害に対する寄与因子の可能性があるものであってもよい。対象が多数の異なる疾患または障害を有する人々であり、属性がその疾患または障害に対する寄与因子の可能性があるものであってもよい。対象が疾患または障害に対する寄与因子の可能性があるものを含み、属性がそれらの因子を持つ、または持たない人々であってもよく、この場合には本方法はその疾患または障害に対する実質的に等価なリスクを持つ人々の群を関連づける。

対象が時間刻みであり、属性がシステムが故障する前の時間刻みでのシステム内の要素の状態を含んでもよく、この場合には本方法はシステムの故障を潜在的に引き起こしうる要素の状態を関連づける。

第1の局面において、r_1はすべての反復に関して同一でありうる。

いずれの局面においても、提供される本方法は、システム状態が選択された時間量にわたる状態変数の値によって提示されるシステム状態間の移行のデータベースをまず作成した上で、各状態から状態への移行の集合がM個の対象のものに対応し、このため各状態変数がある属性に対応するようなデータセットとして全体的または部分的にデータベースを提示するという段階をさらに含みうる。

いずれの局面においても、提供される本方法は、第１に、選択された時間量にわたる状態および作用のデータベースの作成、および各状態／作用／状態の３つ組がM個の対象の一つに対応し、このため各状態変数または作用のタイプがある属性

に対応するようなデータセットとしての全体的または部分的なデータベースの提示、の段階をさらに含みうる。

第19の局面において、本発明は、対象対属性という行列の形式で表現された多数の属性を有する対象のデータセットとともに用いるための一致検出法であって、

・各反復で行列のサンプリングされた部分集合が各対象に関して属性の同じ部分集合を有する、あらかじめ決定された数の反復にわたる行列の部分集合のサンプリング、

・一致が行列のサンプリングされた部分集合における1つまたは複数の対象における複数の属性値の同時出現であって、複数の属性値が各出現に関して同一であって、行列のサンプリングされた各部分集合における一致の検出および数の記録が他の部分集合における一致のサンプリング、検出および数の記録の前、同時または後に行われる、行列のサンプリングされた各部分集合における一致の検出および数の記録、

・サンプリング、検出および記録の前、同時または後に行われる、関心対象の各一致に関する期待数の決定、

・関心対象の各一致に関する一致の観測数と一致の期待数との比較、およびこの比較による、一致に関する複数の属性の相関の程度の決定、ならびに

・k項数の相関属性が、それに関する相関の程度がそれぞれにあらかじめ決定された閾値を上回る複数の属性である、k項数の相関属性の集合の報告

の段階を含む方法を提供する。

第1の局面において、数値的相関値をk項数の相関属性の集合とともに報告することもできる。

図面の簡単な説明

本発明のより良い理解のため、およびそれがいかに実行されうるかをより明瞭に示すために、ここではその例として、本発明の好ましい態様を示す以下の添付の図面に関して言及する。

図1は、部分集合の演算の下に格子として配置された、N＝6の対象を有する集合の幕集合を示したものであり、その幕集合によるすべての可能なk項数の列を示し

ている。

図1aは、図1により示された（黒の四角形）または省かれた（白の四角形）すべての格子結節点の相対的位置を示すものである。

図2は、図1の幕集合に関してn＝1，2，…，6のすべてのサイズのn-グラム（n-gram）を示すものである。

図2aは、図2により示されたまたは省かれたすべての格子結節点の相対的位置を、項の部分集合を強調して示したものである。

図3は、格子の底面からみて第3層の分析に対応する、図1の幕集合に関するすべての可能な対相関（pairwise correlation）を示すものである。これは例えば、蛋白質およびRNA配列ファミリーにおける残基間相関に対してとられる便法である。もう1つの例において、この図は、消費者によって一緒に購入される傾向のある販売品目のすべての対を簡単に見いだす方法でとられるアプローチを示す。

図3aは、図1の幕集合のうち図3に関連する相関を図示したものである。

図4は、図1の幕集合の対象の変数の区分を示したものである。分割は配列ファミリーまたはその他の整列化したデータセットの、1つの特殊で重要な種類の成分モデル（componential model）である。成分モデルにおいて、N_Y個の固有なy_i変数の集合が見出され、N個の観測可能な変数c_iのより大きな集合が「生成」または「説明」できる。分割モデルでは、$N_Y \leqq N$であり、各c_iは厳密に1つのy_iによって生成され、典型的には$N_Y < N$である。1つの固有値に対応する顕在変数は一種の小集団を形成し、おそらく互いには高度に相関し、その小集団以外の変数とは相対的に相関しないと考えられる。図4では、顕在変数は3つの小集団に分けられ

る：(C_1)(C_2、C_5、C_6）および（C_3、C_4）。

図4aは、図1の幕集合のうち図4の区分を図示したものである。

図5は、本発明の1つの態様による、データセットのサンプリングの3回の反復を示したものである。

図5Aは、図5のサンプリングの3回の反復を、注釈とともに示したものである。

図6は、好ましい態様のプログラム法の全体的な流れ図である。

図7は、図6のプログラム法を実行するシステムの模式図である。

図8は、製品の製造のための工程を制御するように適合化された図６のプログラ

ム法の全体的な流れ図である。

図9は、図8の適合化されたプログラム法を実行するシステムの模式図である。

図10は、規則に基づくシステムのための規則を生成し、次いで製品を製造するように適合化された図6のプログラム法の全体的な流れ図である。

図11は、図10の適合化されたプログラム法を実行するシステムの模式図である。

図12は、製品を製造するための工程を制御するために用いられる規則を生成するように適合化された図6のプログラム法の全体的な流れ図である。

図13は、図12の適合化されたプログラム法を実行するシステムの模式図である。

図14は、好ましい態様のハードウエア実装のノードの図である。

図15は、配列の一致が保存された物理的または構造的な関連を示す可能性のある、図15aの標本3D構造の任意の配列に関する残基の図である。

図15aは、標本の蛋白質に関する3D構造の図である。

図16は、本明細書に記載された方法を用いうる、3次構造子測における諸段階の図である。

発明の実施の形態

上に詳述した通り、本明細書に記載される基本的方法では以下の段階を用いる

・対象がある属性を有する場合にその属性が対象において出現すると呼ばれる、

N_A個の変数（「属性」）を項とするM個の対象の集合の表示、

・あらかじめ決定された数の反復における各反復に関する、M個の対象からのr_i個の部分集合のサンプリング、

・一致がサンプリングされた部分集合におけるr_i個の対象のうち同じh_i個における$1 \leqq k \leqq N_A$個の属性の同時出現であって、$0 \leqq h_i \leqq r_i$である、サンプリングされた対象の部分集合のそれぞれにおけるk個の属性の集合間での一致の検出および記録、

・決定がサンプリングおよび収集の前、同時またはサンプリングおよび収集の後に行われる、上記のk個の属性の任意の集合ならびにあらかじめ決定された数のサ

ンプリングおよび一致計数の反復に関する一致の期待数の決定、

・k個の属性の任意の集合ならびにサンプリングおよび一致計数の反復回数に関する一致の観測値と期待数との比較、ならびにこの比較によるk個の属性の集合に関する相関（または結合または依存）の程度の決定、ならびに

・k項数の相関属性が、選択された相関の程度に関してあらかじめ決定された閾値を上回る値を持つことがこの過程によって決定されたN_A個の属性のkの集合である、k項数の相関属性の集合の報告。

1つの代替的な基本的方法は以下の段階を含みうる：

・各反復でデータセットのサンプリングされた部分集合が各対象に関して属性の同じ部分集合を有する、あらかじめ決定された数の反復にわたるデータセットの部分集合のサンプリング、

・一致がデータセットのサンプリングされた部分集合における1つまたは複数の対象における複数の属性値の同時出現であって、複数の属性値が各出現に関して同一であって、データセットのサンプリングされた各部分集合における一致の検出および数の記録が他の部分集合における一致のサンプリング、検出および数の記録の前、同時または後に行われる、データセットのサンプリングされた各部分集合における一致の検出および数の記録、

・サンプリング、検出および記録の前、同時または後に行われる、関心対象の各

一致に関する期待数の決定、

・関心対象の各一致に関する一致の観測数と一致の期待数との比較、およびこの比較による、一致に関する複数の属性の相関の程度の決定、ならびに

・k項数の相関属性が、それに関する相関の程度がそれぞれにあらかじめ決定された閾値を上回る複数の属性である、k項数の相関属性の集合の報告。

本明細書に記載の形態は、上記の基本的方法に対する拡張を提供するものであり、同様の原理を用いる。本明細書に記載される1つの応用の原理を他のものに対して適宜適用してもよい。したがって、応用のすべての構成要素の記載を必ずしもそれぞれの応用に関して繰り返すとは限らない。

好ましい態様においては、プログラミングおよび解釈を簡略化するために、対象が行であって属性が列である行列を用いることが好ましい。しかし、これは厳密に必要とはされず、いずれの態様も、データセットの部分集合を直接サンプリングすることにより、行列の形式で表現されていない対象および属性のデータセットを用いうる。当業者には周知の通り、あらゆるリレーショナルデータベースは2次元行列形式に容易に変換しうる。

本明細書に記載される態様は、多くの異なる標本またはデータセットの他の部分集合の全体にわたってr個の標本のそれぞれに関する一致の検出、記録および計数の段階を同時に実施しうることから、並列処理に特によく向いている。

対象を記述する特徴または変数のそれぞれは数的でも質的でもよい。質的であれば、何らかの数zのレベルまたは質に関して記載された特徴または変数を、z個の可能な値または状態を有する数的変数に変換しうる。z個の可能な値または状態を有する数的変数はz個の二値変数に変換することができ、これは属性と呼ばれる。連続的範囲の可能な値またはレベルを有する数的変数または特徴は、z個の可能な値または状態を有する1つの変数に変換する、またはそれによって表現することができ、このためz個の二値属性の集合に変換する、またはそれによって表現することもできる。

より正式には、本発明者らに、離散値をとるN個の変数v_jのそれぞれに関して、それぞれが特定の値$a_{ij} \in A_j$によって特徴づけられるM個の対象O_1, O_2, …, O_M

からなるデータベースを与えられたと仮定する。特定の変数に関する特定の値は$a_i@v_j$と表記する。連続値をとる変数で始めた場合には、いくつかの既知の方法のいずれかを用いてそれらを離散変数に量子化することができる。また、本発明者らは、多くの応用において、すべての変数に関して可能な値に同一のアルファベットAを用いる。各対象はデータベース中の特定のレコードであってもよく、ランダムな源からの標本であってもよい。

最初のN個の変数が二値的でない場合には、それらをN_A個の属性の集合に変換することができる。例えば、補遺「B」に添付した入力一覧において、各アミノ酸の位置は、アルファベットの文字の部分集合によって表現される20種の天然にみられるアミノ酸に対応する20通りの可能性がある変数である。この変数を二値属性に変換するには、各変数は「A」または「Aでない」、「B」または「Bでない」などのように2つの状態のうち1つをとる20個の異なる属性となる。この種の変数を

表現するための1つの態様は、補遺「A」のソースコード一覧に含まれている。データを属性として表現するためのその他の技法も用いうる。

この説明で述べられた原理を、より高次の計算機で用いる3項属性などのより高次の属性に拡張することもできる。本明細書で用いる二値の例は最も実施が容易なものである。

この状況は、各行が対象を表し各列が属性を表す表によって表現することができ、このため、ここで各表の成分a_{ij}はi番目の対象のj番目の変数がa_{ij}と表記される値を有するという事実を表す。本発明者らは、c_j（「行列j」に関して）および属性を$a_i@c_j$と表記することもできる。

例えば、6行（対象）および6列（変数）からなるこの小さな行列を考える。

行1	行2	行3	行4	行5	行6
A	B	C	D	E	F
W	U	C	V	E	G
Z	L	C	M	W	M
V	U	C	V	A	G

A　B　C　D　Z　Z

W　L　C　M　E　Z

↑　↑

対象番号1は変数1に関して「A」、変数2に関して「B」、変数3に関して「C」といった値をとる。いくつかの用途には、例えば変数2および4が相関しているといったことが見いだされることが有用と考えられる。上記の小型の（小さな架空の）行列の例では、対象がB@2を有する時には常にD@4も有し、対象がL@2を有する時には常にM@4も有し、対象がU@2を有する時には常にV@4も有していることから、この相関は妥当と思われる。属性番号3は変化せず、あらゆる対象が属性C@3をとっているため、これは他のどの変数とも興味深い様式では相関しない。

あるデータの行列が与えられたとして、本発明者らはさらに、すべての次数k＝1，2，…，N_Aに関してそれぞれの可能なk項数の属性に関する確率を特定する何らかの「真の」基礎確立分布q( )が存在すると仮定する。例えば、k＝1については、本発明者らは$q(C_j):A_j \rightarrow [0, 1]$を有し、本発明者らは何らかのデータセット

に関してq(B@2)＝0.33を有すると考えられる。分布が、例えばq(B@2，F@6)＝0.166というように、より高次の確率を特定することもある。提起されたこの特定の問題には、分布q( )もしくはその少なくとも一部の推定または概算の問題が内在する。

問題は、k＝2，…，N_Aに関して、相関があらかじめ決定された何らかの値よりも高い、いくつかまたはすべてのｋ項数の行（c_{j1}，c_{j2}，…，c_{jk}）を見いだすことである。例えば、M×Nの値の表が与えられたとして、何らかの実数ρ_kに関して$D(q(v_{j1}, v_{j2}, \cdots, v_{jk} \mid \Pi_{i=1 \cdots k} q(v_{j1})) > \rho_k$であるようなk項数の行添数（$j_1$，$j_2$，…，$j_k$）の一覧を返す手順を必要とすることが考えられる。ここで$D(p_1 \mid p_2)$はカルバック発散測度であり、この場合には、行変数の全体にわたり観測された値の分布とすべての行変数が統計的に独立している分布との間の差を推定している。カルバック測度は、この種の問題に適用しうる多くの考えられる相関または関連性の指標のうちの1つに過ぎない。

本発明者らの目的のために、本発明者らは、統計的独立性からの偏差に関する

相関を検討した。独立変数が存在するという基礎仮説が真であるとして、データベースの調査における何らかの事象の観測発生数と期待される数とを比較することができる。すなわち、問題は以下の通りである：値の表が与えられたとして、（$a_{i1}@c_{i1}$，$a_{i2}@c_{i2}$，…，$a_{ik}@c_{ik}$）の何らかの観測挙動、何らかの実数閾値$\theta_i \in [0, 1]$およびその推定または仮説検定の方法の基礎となる何らかのモデルに対し、$k=2 \cdots N_A$のすべてに関して、P(Observed($a_{i1}@c_{i1}$，$a_{i2}@c_{i2}$，…，$a_{ik}@c_{ik}$)|IndePendent(c_{i1}，c_{i2}，…，c_{ik})，Model)$< \theta$であるようなすべてのk項数の属性（$a_{i1}@c_{i1}$，$a_{i2}@c_{i2}$，…，$a_{ik}@c_{ik}$）の一覧を返す。

標本抽出のサブプロセスは無作為抽出でよく、無作為的である場合には均一分布を含む、対象に関する多数の可能な確率分布のいずれかの支配下にありうる。同様に、本方法の演算の間に抽出されたT個の標本のそれぞれの間、および1つの標本内で抽出されたr個の対象のそれぞれの間の統計的独立性または依存性に対して拘束があってもよい。

好ましい態様の利点の例

上記および以下にさらに記載される一致検出法および装置の比較優位性が最も明らかな、多くの多様な応用領域において生じる問題が少なくとも1種類ある。このような問題は以下を特徴とする：

１．多数の属性（本発明者らの表現では行）、

２．互いに高い相関がある属性の何らかの数の小集団であって、このような小集団の各メンバー属性がそれ自身の小集団以外の属性とは相対的に相関しないような小集団がデータセット中に存在する可能性がある、ならびに

３．このような属性小集団の正確な数、幅（k項相関およびk次の特徴などにおけるk）および位置に関する予備知識がない。

本発明者らが知るすべての他の手順は、発見可能なk項数の幅kに対する事前の制限を設定するか、または連続的もしくは並列的な網羅的な検索をすべてもしくはほぼすべての可能なk項数の属性にわたって実行するかのいずれかである。これをより簡略化するために、好ましい態様の方法では、44項相関を見いだすために要する計算時間およびメモリは、同じ極めて高次元のデータセットにおいて2

項相関を見いだすために要するものとほぼ同一である。これに対して、大部分の従来法では、44次特徴の発見を除外するか、さもなければそれを見いだすために何桁も大きい時間または空間を割り当てる必要がある。

好ましい態様の応用の例

極めて大量なデータセットの製作者は、タスクの計算上の複雑性、および大部分の高次項に関して統計的に有意な推定を裏づけるために必要なデータの不足という両面により、完全に高次な確率モデルによる高度の計算を行うという試みを阻まれてきた。

好ましい態様では、データベースモデルを構築するために、高次確率の部分集合のみを計算し、高次特徴（「HOF」）を限定的に選択した上で抽出する。本明細書に記載される相関検出法を用いて高次特徴の集合をあらかじめ選択すること、ならびに最も有意なもの（統計的、および用途特異的な基準に関して）を既存の統計的な、規則に基づく、神経回路網による、または文法に基づく方法に基づいたモデルに基づく分類機（classifier）および予測機（predictor）に組み入れることにより、限定的な計算資源からの効率的な利用が可能となる。あらかじめ選択されたHOFの集合は、このようなシステムのための規則を作るために用いること

ができる。例えば、ある会社が特許出願を提出しようとしている際に発明者からの譲渡証を提出すべきかどうかを決定するために、本明細書で詳述した方法を用いてデータセットを分析することもできる。次にこの規則は、会社が特許出願を提出しようと決定する際に常に譲渡を生成するためのシステムに用いられる。規則に基づく多くのネットワークは、本明細書に記載される方法を用いる予備処理によって利益を得ると考えられ、例えば、1992年10月27に発行された米国特許第5,159,662号に記載されたグレーディ（Grady）らのコンピュータに基づく網パターンマッチング回路網を構築するためのシステムおよび方法（the System and Method for Building a Computer-Based Rete Pattern Matching Network）、1992年6月2日に発行された米国特許第5,119,470号に記載されたハイランド（Highland）らの推論エンジン、ならびに1993年1月12日に発行された来国特許第5,179,6

32号に記載されたマツイ（Matsui）らの双方向推論のための迅速法（the Fast Method for a Bidirectional Inference）などを参照されたい。

または、見いだされたHOFを、例えば、距離幾何学的または経験的に推定された協同性およびフォールディングのパターンに基づく既存の方法に導入した場合の蛋白質構造の予測または決定において、または相関する製品の販売情報に基づく販売計画において、製品の作製のために直接的に用いることもできる。

以下では、ロスアラモス（Los Alamos）HIVデータベースを用いる本明細書に記載された原理の実践について説明する。特にこれらの原理は、ヒト免疫不全ウイルス（HIV）のエンベロープ蛋白質のV3ループの検討に適用された。生化学および分子生物学の全般において、蛋白質の特定の残基に共変がみられることは、機能的、生理的役割を持つ蛋白質の領域を特徴づける構造モチーフの存在を示す可能性が高い。

エンベロープ蛋白質はウイルス粒子の周囲を取り巻く脂質膜中に部分的に包埋され、脂質から外部に突出する。感染時にHIV粒子の脂質が宿主細胞の膜と融合すると、エンベロープ蛋白質も感染細胞の膜から突出することがある。V3ループの配列は様々なウイルス単離株の間で非常に異なることから、V3のVは「可変な」という意味を表す。

以前、ロスアラモスグループのB.T.M.コルバー（Korber）、R.M.ファーバー（Farber)、D.H.ウォルパート（Wolpert）およびA.S.ラペデス（Lapedes）は、本明細書に参照として組み入れられる「HIV-1のV3ループにおける共変：情報理論による分析（Covariations in the V3 loop of HIV-1:An information-theoretic analysis）」、Proc. Nat. Acad. Sci. U.S.A. 90（1993）において、HIV-1エンベロープ蛋白質のV3ループの特定の残基における2項共変変異を述べている。本原理の実践により、ロスアラモスグループの結果の一部は確認されたが、さらに、その他の高度に共変する残基の群の発見も可能であった。ロスアラモスグループは対共変を見いだしたのみであったが、本発明者らは本明細書においてk＞2であるk項残基の共変を記載する。すなわち、本発明者らはHIVエンベロープ蛋白質のこれまで認識されていなかったモチーフを同定した。

特定の試行に関して、入力は657種の異なるウイルス単離株由来のV3領域のそれぞれのアミノ酸配列からなり、それらは補遺「B」に示されている、入力に関して用いたソースコードは補遺「A」および「D」にそれぞれ「File coinc.pl」および「File probsort.pl」の名称で示されている。出力は補遺「C」に示されている。

以下の別項に詳述する表C.1からC.9までを参照すると、6回の別々の試行の結果が示されている。パラメーターの値はそれぞれの凡例に表記した通りである。各表では、結果は最も有意な相関が最初になるように統計的有意性の順に示され、標準的な1文字アミノ酸コードを用いている。したがって、表C.6を参照すると、観測された最も有意な一致は残基18のアラニン（A）、残基31のグルタミン（Q）および残基33のヒスチジン（H）の出現である。これは、引用したページで示された他の一致と同様に、これらの残基を含むHIV-1 V3ループの構造モチーフが同定されたことを意味する。

A18／Q31／H33の特定の例を続けると、V3の構造モチーフはおそらくウイルス粒子の外側に存在するこれらの残基を含み、V3ループのその領域は特定の構造モチーフを必要とする特定の機能を果たす可能性が高い。このため、この構造モチーフはその機能を保持するために変異後も保存される必要があると考えられる。この推論は本明細書で同定される他の一致にも拡張される。

HIVの特定の保存された構造モチーフが同定されることにはいくつかの用途がある。

当技術分野で知られた技法を用いることにより、このモチーフを具現化したペプチドを抗体として用いうると考えられる。したがって、ワクチンを調製することができる。このモチーフを具現化したペプチドは、例えばマニアティス（Maniatis)ら、分子クローニング：実験室マニュアル（Molecular Cloning:A Laboratory Manual）、Cold Spring Harbor Laboratory、Cold Spring Harbor、NY（1982）およびサムブルック（Sambrook）ら、分子クローニング：実験室マニュアル（Molecular Cloning:A Laboratory Manual）（第2版）、Cold Spring Harbor Laboratory、Cold Spring Harbor、NY（1989）などに一般に記載されている既知

の組換え法を用いて作製してもよい。または、ペプチドまたは疑似ペプチドを標準的な化学的技法を用いて化学合成してもよい。ペプチドまたは疑似ペプチドに対するモノクローナル抗体を、例えばハーロウ（Harlow, E）およびレーン（Lane, D.）、抗体：実験室マニュアル（Antibodies:A Laboratory Manual）、Cold Spring Harbor Laboratory、Cold Spring Harbor、NY（1988）に記載されたものなどの標準的な方法を用いて作製してもよい。新規な構造モチーフに対する特異的親和性を有するこのようなモノクローナル抗体の断片、例えばF_{ab}断片を作製することもできると考えられる。

もう1つの態様では、本発明に従って同定された構造モチーフと相互作用するリガンドを作製することができる。すなわち、このリガンドは適切な実体をもち、モチーフの対応する残基または部分と相互作用するように互いに位置する化学成分（chemical moiety）を有することを特徴とすると考えられる。いくつかの態様において、リガンドはそのモチーフと結合することによってその領域の機能を妨げる作用物質、例えば薬物でありうる。このため、リガンドは潜在的な治療的有用性を侮えたHIV拮抗薬になると考えられる。または、同定されたモチーフを含む特定のV3領域とリガンドを結合させ、診断的有用性を得ることも可能である。このような診断的有用性はエクスビボ的でありうる。診断的有用性を備えたリガンド（例えば抗体）は、比色反応に用いるための蛍光または酵素複合体などの標識を含むこともできる。蛍光標識されたウイルスまたはウイルス感染細胞は蛍光顕微鏡またはFACS（蛍光標示式細胞分取器）を用いて可視化または計数しうる。

本発明に従って同定された構造モチーフと結合するリガンドの設計および同定の方法も、本発明によって提供される。

したがって、1つの態様において、本発明は、アミノ酸残基A18／Q31／H33を含む構造モチーフを含むヒト免疫不全ウイルス（HIV）のエンベロープ蛋白質と結合するためのリガンドを提供する。このリガンドはそのモチーフと結合しうる少なくとも1つの官能基を含む。1つの好ましい態様において、リガンドは、残基18との結合能を有していて前記リガンド中の残基18との結合のための有効部分に存

在する少なくとも1つの官能基、残基31との結合能を有していて前記リガンド中の残基31との結合のための有効部分に存在する少なくとも1つの官能基、および残基33との結合能を有していて前記リガンド中の残基33との結合のための有効部分に存在する少なくとも1つの官能基を含む。

もう1つの態様において、本発明は、ヒト免疫不全ウイルス（HIV）のエンベロープ蛋白質の構造モチーフと結合するリガンドを設計する方法を提供する。本方法は、HIV-1エンベロープタンパク質のV3ループ内にA18、Q31およびH33の空間座標を有するテンプレートの提供、および空間的拘束を備えた有効なアルゴリズムを用いて、モチーフと結合する少なくとも1つの有効な官能基を含むような化学的リガンドを計算的に展開することを含む。1つの好ましい態様において、リガンドは、残基18との結合能を有していて前記リガンド中の残基18との結合のための有効部分に存在する少なくとも1つの官能基、残基31との結合能を有していて前記リガンド中の残基31との結合のための有効部分に存在する少なくとも1つの官能基、および残基33との結合能を有していて前記リガンド中の残基33との結合のための有効部分に存在する少なくとも1つの官能基を含む。

もう1つの態様において、本発明は、ヒト免疫不全ウイルス（HIV）のエンベロープ蛋白質の構造モチーフと結合するリガンドを同定する方法を提供する。本方法は、HIV-1エンベロープタンパク質のV3ループ内にA18、Q31およびH33の空間座標を有するテンプレートの提供、分子の構造および配向性を含むデータベースの提供、ならびに前記分子がモチーフと相互作用するように互いに空間的に配置された有効な成分を含むかどうかを決定するためのスクリーニングを含む。1つの好ましい態様において、分子の第1の成分は残基18と相互作用し、分子の第2の成分は残基31と相互作用し、分子の第3の成分は残基33と相互作用する。

本明細書に記載される原理は、本明細書に記載される他の共変するk項数、すなわち共変するV3ループの両方の残基、および共変する特定の残基の特定のアミノ酸に関する、抗原およびワクチンを含む同様のそれぞれの態様を包含する。

本発明の方法は、高次特徴を検出するための「高域フィルター」とみなしうる。このようなHOFは、データベースのモデル化、機械学習、ならびに知覚および

パターン認識において重要な役割を果たす。データベースのマイニングおよびモデル化の文脈において、これらの特徴を見いだすための手順は以下を含むいくつかの主要な役割のうちのいずれかに役立つと考えられる。

１．大量の複雑なデータセットの予備処理：ギブズモデル、隠れマルコフモデルおよびEM、マッカイ（MacKay）の密度ネットワーク（density network）、および神経回路網分野での関連した要因学習法（factorial learning）を含む最も優れたモデル化法の多くは、本明細書に記載される原理を実行することによって提供されるものなどの、データベース中の相関すると思われる変数を見いだす迅速な予備処理手順を先行することにより、網羅的検索またはパラメーター空間の組み合わせによる爆発的増加を伴わずに高次相互作用を捕捉する点において大きな助けを得ると考えられる。

２．大量の複雑なデータセットの視覚的表示：最も単純なグラフィカルディスプレイインターフェースと組み合わせた場合ですら、本発明者らのものなどの手順により、ユーザーが高次元データにおける最も可能性の高いと思われる興味深い高次特徴を迅速に（少数のr個の標本で）観察することが可能になる。

３．予備調整（pre-conditioning）および冗長性の排障：ここまでは、本発明者らはモデルの構築に用いるための属性間相関を見いだす有用性を強調してきたが、多くの最適化、学習およびデータ適合化の用途においては、主成分分析法（PCA）などの多数の部分空間法のいずれかにより、変数の間の相関を発見および排除する必要がある。

プログラム可能なデジタルコンピュータを用いる1つの態様

デジタルコンピュータの態様のための構成要素

データ行列、標本抽出および一致　N_A個の属性の固定集合のそれぞれに関して、それぞれが「はい」（1によって表現される）または「いいえ」（0によって表現される）のいずれかの値をとるM個の対象の集合が与えられた場合、入力データセットを$M \times N_A$の値の表として配列することができ、これを本発明者らはデータ行列（data matrix）または単に行列と呼ぶことにするが、この行列は、以下に説明するシステム／プロセスの機能的部分を構成する部分行列および関連ベク

トルとともに、プログラム可能なコンピュータの内部のメモリ位置に保存される。この表現において、行列の行は対象に対応し、列は属性に対応する。この行列をV_{ij}と表記し、この二次元表の各要素を$v_{ij} \in \{0, 1\}$によって表記しうるが、ここでiはi番目の対象（行）o_iを意味し、jはj番目の属性（列）a_jを意味する。これを記載する目的で、対象の集合を$O=o_1, o_2, \cdots, o_M$としてリスト化してもよく、属性の集合を$A=a_1, a_2, \cdots, a_{NA}$としてリスト化してもよい。

図5Aは、好ましい態様のプログラム法の説明に関して以下により詳細に考察される、図5に例示された実施例に対して適用されたこれらの項を例示したものである。

$a_{ij}=1$であれば、特定の属性a_jが特定の対象（列）iにおいて出現するということができる。

$1 \leqq m \leqq M$個の対象（列）5の順序付きのリストが与えられた場合、ある属性a_jに対する出現ベクトル2は、与えられた対象のリストにおけるg番目の対象において属性a_jが出現する場合、しかもその場合のみにg番目のビットが1であるような二値ベクトルまたは長さmの列（string）として定義しうる。出現ベクトル2は、いくつかの対象の集合にわたる属性の出現パターンを単純に表現したものであり、例えば、M個の対象すべての集合、または以下に説明する実施例に対応する対象の集合である。

サンプルr、例えば図５Ａにおける参照する数字４によって同定される３列であるが、ある確率分布からランダムに導き出された記録Mからなるrのセットである。ある好ましい態様では、サンプル中の列は、一定の分布から独立して導き出されるものと考えられる。

各あらかじめ決定された反復数の中で一回という系で、サンプルrのサンプル４の製図を行った。ある好ましい態様では、全部でＴ回の反復によって導き出されたサンプルは、一定の分布から独立に導き出されるものと考えられる。

いくつかの好ましい態様では、平行したコンピュータ計算の実施態様において、異なる連続反復サンプリング、および/またはさまざまな処理ノードによって処理されたデータセットの異なったサブセットに対して、さまざまなr値が用

いられる。このような場合に、i番目の反復について、またはi番目のサンプルにおいて、対象となるザンプルの番号はrであるという。異なったサンプルサイズを用いることの長所には、その方法を一通り行なう中で、どのr値が最適であるかがはっきりしないときに、さまざまなr値を試すことができること；また、さまざまな処理ノードの中で、さまざまなプロセッサーサイズ/速度、およびメモリサイズを最適に使用するために、平行コンピュータ計算における異なった処理ノードに対して、異なったr値を選び出すことができることなどがある。その方法を一通り行なう問中、同一のr値を一つだけ用いることの長所は、プログラムのコードが単純になるということが僅かな利点となるだけである。

一致集合（set）、またはCsetは、対象（列）５の何らかの集合内の$1 \leqq k \leqq NA$属性（行）1の結合アピアランスを含むパターンとして定義することができる。すなわち、何らかの一つもしくは複数の列５を考慮に入れると、$a_{j1}, a_{j2}, \cdots$およびa_{jk}が全て一つもしくは複数の所与の列に生じる場合、cset $a_{j1}, a_{j2}, \cdots, a_{jk}$が存在する。たとえば、図5Aに示した式３によって決められる要素A@c1、B@c2、D@c4は、セット（ｃセット）に一致する。

コンピューターメモリーはｃセットテーブルと呼ばれるデータ構造を記録し、それは前記プロセスにおける一つあるいはそれ以上の繰り返しが出現する各ｃセットの同定、および出現数の連続を意味する。ｃセットの同定は、ｃセットを構成する属性（列）のリストであり；出現数は、このプロセスにおける特定の繰り返しが見つかるまで、あるいは全ての繰り返しの終わりまでのｃセットの出現数の数に対応する数である。他の態様においては、コンピューターのメモリーに記録されたハッシュ･テーブルとしてｃセット・テーブルが満たされる。

あるr標本のためのcsetは、標本中のrデータアイテムにわたってその発生（「］」により表示される）および非発生が（「0」）、二進法によりコードされた記録である、特定の発生率のベクトルを有する。従って、k属性の集合に対応するcsetは、関連する発生率のベクトルを有する可能性があり、およびそれぞれの属性

は、関連する発生率のベクトルを有する可能性がある。

あるr標本において、あるcset $\alpha=(\alpha_{i1},\ldots,\alpha_{ik})$に対して$\alpha_{i1}$が、r記録中のhにおいて現れるとき、...、およびα_{ik}がr記録中のhにおいて現れるとき、サイズhのマッチ（または発生率）が生じているといい、それらはr記録中の同じhにおいて確実に現れる（図５Ａを参照）。

一致の観測数　一致を観測し、対応するcsetを「ビンニング（binning）法」により保存またはアップデートする。各反復(iteration)において、属性を二値列化（binned）し、分離されたサブセットの中に、その時の反復におけるそれらの結合（incidence）ベクター２割るｒ－試料４に従い配置した。ここに記載された本発明のマトリックスに基づく態様において、これらのベクターは、2^rアドレススペースの、非常にまばらなサブセット中に配置（addresses）するｒ－ビットのように振舞う(図５および5A参照)。

１回の二値列（bin）における全ての属性は1つのcsetを構成する。このcsetは記録され、仮に特定のcsetが過去の反復において起こったものである場合は、生起の回数はアップデートされ、仮に過去に起こったことのない場合は、そのためにcset表のエントリーを作成し、その生起回数がアップデートされる。ここに記載された態様では、系は数ｈを保存する：この反復および各反復において、$0\leq h\leq r$の生起。反復の数Ｔの特定が完了した後、cset表は、観察された全てのcset、および、$\Sigma^{l}_{r=1}\ h_l(\alpha)$で表される、各cset$\alpha$に対する「観察された生起」の総数のリストを含む。ここで、$h_l(\alpha)$は、ｉ回目の反復におけるα_iを含むｋ属性に対する結合生起の数を表す。

期待された関数計算　期待された関数計算は、数学的なものであり、コンピュータープログラムあるいはサブルーチンとして、または電子工学あるいは光学式回路として働く。それは、$a_{j1},a_{j2},\cdots,a_{jk}$という特質及びTの数の１セットを導き、

γ試料の描写及び観察上一致したTの反復の際に、特質の1セットに対し、期待された一致数に符合する。

本発明の1つの特定の態様では、多項分布から関数$f_{match}(\alpha,h,r)$が得られる：

$$f_{match}(\alpha, h, r) = \left(\frac{r!}{h!(r-h)!}\right) p(a_{i1}, \ldots, a_{ik})^h p(\bar{a}_{i1}, \ldots, \bar{a}_{ik})^{r-h},$$

この式は、1つのr標本において、すべてが同一のh行に生じる、a_{i1} の厳密にh回の出現、a_{i2} の厳密にh回の出現、…およびa_{ik} の厳密にh回の出現が見いだされる確率の推定値を与える。

（標準的な多項式において2を除くすべての多数のp()因子はゼロ冪指数に伴って消失するため、この関数の定義は単純な形式を有する）。

可能性のあるcsetを構成するk個の属性に関する大きさhの合致の確率は同時確率$p(a_{i1}, \cdots, a_{ik})$に関して定義されており、期待計数関数数はこれらの同時確率に関して特定の推定値を用いる必要がある。この好ましい態様において、同時確率の推定値は個々の属性の間に独立性があるとの仮説を取り入れている。したがって、上記で与えた定義式において本発明者らは$P(a_{i1}, \cdots, a_{ik})$の代わりに

$\Pi_{i=1}^{k} p(a_{ii})$ を、 $p(\bar{a}_{i1}, \ldots, \bar{a}_{ik})$ の代わりに

$\Pi_{i=1}^{k} (1-p(a_{ii}))$ を用いる。

<u>仮説検定関数および相関速度</u> 仮説検定とは、コンピュータプログラムもしくはサブルーチンとして、または特殊な目的の電子的および／または光学的ハードウエアにおいて履行される数学的手順であり、k個の属性の特定の集合に関してそれぞれが一致の期待数および観測数を表す一対の数字H_{exp} およびH_{obs} を用い、k個の属性間の相関の推定値を表す数Cを生成する。

いくつかの好ましい態様では、以下に説明する通り、尾部確率に関するチャーノフ境界により仮説検定関数が提供される。

無作為変数X_iが各反復に関して値h_iをとるとし、 $X = \Sigma_{i=1}^{r} X_i$ とし、ここで$0 \leqq X \leqq T \cdot r$とする。チャーノフ-ヘフディング境界［8］の方法により以下の理論が与えられる：

$X = X_1 + X_2 + \cdots X_n$をn個の独立無作為変数sの合計とし、ここで実数$l_i$（「下位値」）および$u_i$（「上位値」）に関して$l_i \leqq X_i \leqq u_i$とする。

すると、

$$P[X - E[X] > \delta] \leqq \exp(-2\delta^2 / \Sigma_i (u_i - l_i)^2) \quad (1)$$

本発明者らの目的のために、本発明者らはすべてのi＝1，2，…，Tに関してn＝Tおよびl_i=0およびu_i=r_iと設定し、本発明者はそれによって以下を得る。

$$P[X - E[X] > \delta] \leq \exp\left(\frac{-2\delta^2}{\Sigma_i r_i^2}\right) \qquad (2)$$

この数学的関係を用いて、相関値を計算するための有効な手順を規定することができる：

$$Corr(\alpha) = 1 - \exp\left(\frac{-2(H_{obs} - H_{exp})^2}{\Sigma_i r_i^2}\right).$$

標本抽出のあらゆる反復に関して同一の標本数rが用いられる、すなわちすべてのi＝1，2，…，Tに関してr_i＝rである特殊な場合には、上記の式は以下のより単純な形式に変わる。

$$P[X - E[X] > \delta] \leq \exp\left(\frac{-2\delta^2}{Tr^2}\right) \qquad (2a)$$

$$Corr(\alpha) = 1 - \exp\left(\frac{-2(H_{obs} - H_{exp})^2}{Tr^2}\right)$$

ここで相関値は、期待数H_{exp} の基礎にある仮説が真である場合には、r個の標本抽出のT回の反復にわたり観測されたH_{obs} 一致を有する確率を1から差し引いた推定値に対応する。いくつかの好ましい態様に関して、属性間の独立性の仮定が上記のH_{exp} の計算に用いられた場合には、この仮説検定により、独立性からの偏差を推定する各csetに関する相関値が与えられる。すなわち、それにより、csetを構成する属性間の統計的依存性が推定される。

<u>プロセス内部の構成要素の演算</u>

典型的には、表現構成要素はまず本発明の全体的プロセスの内部で動作する。複数の標本抽出の反復がデータの表現に対して実施され、各r-標本に関して、一致の検出および記録がなされる。

標本抽出の反復は連続的に行っても並列的に行ってもよく、または連続的および並列的ステップの何らかの組み合わせで行ってもよい。

プロセス内の任意のステップで、属性の一致集合の一部または全体に関して、一致の期待数の決定が行われる。プロセスのこの構成要素は、すべての一致集合

に対してすべて一度に行っても段階的に行ってもよく、連続的もしくは並列的または何らかの組み合わせで行ってもよい。各一致が検出または保存された時に一致集合（cset）に関してこれを行ってもよく、このような検出または記録の前または後に行ってもよい。

何らかの数の標本抽出の反復を行った後に、記録された一致集合の一部または全体に対して、一致の実際の数と期待数との比較を行うことができる。これはすべてのcsetに関して一度に行ってもよく、またはプロセスの全体を通じて種々の時点でそれらの部分集合に関して行ってもよい。異なるcsetに関するこれらの比較は連続的に行っても並列的に行ってもよく、その何らかの組み合わせで行ってもよい。

何らかの数の標本抽出の反復を行った後に、比較により構成要素の属性間に有意な相関が認められると判定された、記録された一致集合の一部または全体に関して、相関属性の集合を報告してもよい。これはすべてのcsetに関して一度に行ってもよく、またはプロセスの全体を通じて種々の時点でそれらの部分集合に関して行ってもよい。異なるcsetに関するこれらの比較は連続的に行っても並列的に行ってもよく、その何らかの組み合わせで行ってもよい。

好ましい態様のプログラム法の説明

以下には、プログラム可能なデジタルコンピュータに関する1つの可能な態様に対応する、フロッピーディスク、ハードディスク装置、RAMまたはその他の媒体などの適切な媒体上にあるプログラムが疑似コードの形で示される。

図5は、架空の小型データセットに対するこの態様の適用の例を図面で提供している。小型データセットに対するr個の標本抽出（r＝3に関して）の3回の反復が上から下の順に描写されている。各反復に関して、左側の枠はデータセットを表し、その中の枠で囲まれた部分はサンプリングされた行を表す。右側の枠は、属性が重なる二値列（bin）の集合を表す。例えば、1回目の反復では、サンプリングされた3つの行のうち1番目および2番目で、A@1、B@2およびD@4のすべてが生じており、このためそれらはそれぞれ一致ベクトル110を有し、その二進アドレスによって表記された二値列において重なる。単一の属性のみを含む二値列は無

視され、「空白の」二値列は全く生成されない。すべての二値列は各反復の後に消去および除去されるが、重なり（collision）はCsetの全体的データ構造に記録される。

相関属性の集合を見いだすための手順：

```
0. begin
1.    read(MATRIX);
2.    read(R, T);
3.    compute_first_order_marginals(MATRIX);
4.    csets:={};
5.    for_iter=1 to T do
6.    sampled_rows:=rsample(R, MATRIX):
7.    attributes:=get_attributes(sampled_rows);
8.    all_coincidences:=find_all_coincidences(attributes);
9.    for_coincidence in all_coincidences_do
10.   if cset_already_exists(coincidence, csets)
11.   then_update_cset(coincidence, csets);
12.   else add_new_cset(coincidence, csets);
13.   endif
14.   endfor
15.   endfor
16.   for cset in csets do
17.   expected:=compute_expected_match-count(cset);
18.   observed:=get_observed_match_count(cset);
19.   stats:=update_stats(cset, hypoth_test(expected, observed));
20.   endfor
21.   print_final_stats(csets, stats);
22.   end
```

疑似ユードのステップ5から21までは本明細書に記載される基本的方法の段階

、すなわち以下を表す：

・属性の各部分集合が同一である、あらかじめ決定された数の反復に関する行列の部分集合のサンプリング、

・一致がサンプリングされた部分集合における1つの対象における複数の属性値の同時出現であって、各出現に関して複数の属性が同一である、サンプリングされた各部分集合における属性の一致の検出および数の記録、

・サンプリング、検出および記録の前、同時または後に行われる、関心対象の各一致に関する期待数の決定、

・関心対象の各一致に関する一致の観測数と一致の期待数との比較、およびこの比較による、一致に関する複数の属性の相関の程度の決定、ならびに

・k項数の相関属性が、それに関する相関の程度がそれぞれにあらかじめ決定された閾値を上回る複数の属性である、k項数の相関属性の集合の報告。

補遺「B」は、Sun4コンピュータ上でサン（Sun）UNIXオペレーティングシステムにおいて動作させるための、Perl言語で記述された実際のソースコードを含む。補遺「B」のコードリストのための入力データのサンプルを、補遺「C」にHIVエンベロープ蛋白質のV3ループ由来の部分的アミノ酸配列に関して記載している。補遺「C」の入力に関する補遺「B」のコードからの対応する出力を補遺「D」に示している。補遺「D」の出力を生成する目的で、補遺「B」の主コードリストの説明および表現のために、補遺「E」に記載した補助的Perl言語プログラムを用いた。この態様に関する全体的な流れ図は図6に示されており、全体的なブロック図は図7に示されている。結果として得られた報告は相対的にみて体系化されていないアスキー（ascii）データベースとしてフラットファイル中に保存され、それが後に印刷される。それをプリンタに直接送ること、または他のリソースへの報告のためにネットワークを通じて送ることも同じく可能である。

代替的な態様

本発明の代替的な態様の説明は2つの範嗚に分けることができ、以下に別々に説明する：その1つは問題特異的（problem-specific）な可能性のある多くの用途において用いうるようなシステム／プロセスの種々の物理的態様であり、2つ

目は本発明の種々の問題特異的用途による、上記の説明において列挙した構成要素の異なる解釈である。

異なる履行

例えば、プログラム可能なデジタルコンピュータ上のプログラムとしての多くの可能な態様のうち：

本方法は上記に与えられた疑似コードの最も直接的な解釈において完全に連続的に実行することもでき、または本方法は並列的（ベクトルまたはマルチプロセッサ式）または分散型コンピュータシステム上で多くの可能な方式で実行することもできる。一組の計算を、各計算がそれぞれの別々の計算でrすなわち標本数に関して異なる値を用いる点を除いて上記に概要を述べた通りのプログラムのステップ全体を実施する一組の計算を並列的に実行してもよいが、無作為なr-標本抽出のために異なる初期無作為数の開始点（seed）から開始する条件で、それぞれの別々の計算は同一の重要なパラメーター値を有する同一のプログラムステップを実行することもできる。または、それぞれの異なるr-標本を異なるプロセッサ上で実行する別々のプロセスに分けるという条件で、上記に概要を述べたプログラムのステップ全体を1回実行することも可能であると思われ、ここでそれぞれのこのようなプロセスは検出および選択的には記録の段階を含み、全体的なcsetの数が後に連結されて全体的なプロセスおよび全体的なデータ構造になると考えられる。さらに、期待数の計算および期待数と観測数との比較はすべて一度に行っても段階的に行ってもよく、連続的または並列的に行ってもよい。同様に、推定された相関値の報告は、Csetの一部または全体に関して計算の終了時に一度に行っても段階的に行ってもよく、または全体を通じて連続的もしくは並列的に行ってもよい。

有意に相関するk項数の属性（比較、すなわち仮説検定の段階で十分高度に相関すると思われたcset）の報告を含みうる本方法の出力は、言葉によるものでもよく、ならびに／または数字および／もしくは図形によるものでもよい。

標本抽出方式としては、確定的、疑似無作為的または純粋に無作為的なものを含む、多数のものが可能である。疑似無作為的または無作為的であれば、超幾何

的および多項的な標本抽出を含む、多数の無作為標本抽出方式のいずれを用いることもできる。r-標本内部のr個の対象の標本抽出は「置換あり」でも「置換なし」でもよい。次に高いレベルに進んだ場合には、r個の標本それ自体の集合を「置換あり」または「置換なし」で抽出することもできる。

重要な標本抽出パラメーターrに関する異なる選択も可能であり、各標本に関して同じ数rを用いる必要はない。

標本抽出の反復回数Tについては多くの選択が可能である。本発明の方法によって見いだされたk項数の属性に関して推定された相関の程度において望ましい信頼

度を達成するためのTを選択するには多数の数学的方法のいずれを用いることも可能である。または、手順を任意の一定回数にわたり反復して実行した後に結果を印刷または表示すること、または何らかの回数にわたる反復の実行と結果の印刷または表示をインターリーブにより行うことも可能である。

アルゴリズムの処理の間に用いられるCsetデータ構造の表現、保存およびアクセスのためには多くの可能な方式が存在する。Csetデータの保存およびアクセスは、ハッシュ表（hasht able）、k-dツリー、パトリシアツリー（trieとも呼ばれる）および／またはデータの効率的な保存およびアクセスのための当業者に知られた他の方法によって行いうる。いかなるデータ構造が選ばれた場合でも、その構造はレジスタ内、主記憶装置内および／または磁気ディスク、磁気テープもしくは光学的保存媒体などの二次的もしくは外部の保存媒体に物理的に保存されうる。

種々のタイプの汎用計算ハードウエアに対する本方法の態様に代わるものとして、特殊用途の電子的、光学的もしくは電気光学的ハードウエア、または汎用および特殊用途のアーキテクチャーおよび装置の何らかの組み合わせに対する多数の可能な態様も存在する。

例えば、本発明の行列表現を実行するために極めて効率的な特殊用途電子回路（LSIまたはVLSI）を用いることもできるが、これは属性の発生ベクトルが単純な二進ベクトルであるという事実、本発明の1つの考察において以前に説明した

一致「二値列（bin）」が各r個の標本に関して大きさ2^rのメモリ空間に対する「アドレス」に対応するという事実、ならびに無作為数の生成およびサンプリング、Csetデータ構造の高速アクセス保存ならびに期待数推定値の計算および仮説検定および相関推計に用いられる数学的関数の実行のための特殊用途ハードウエアの設計、製造および使用が現在の技術で可能であることによる。

1つの好ましい態様の特殊用途ハードウエアの方法の説明

１．概要

図14を参照すると、上記の特殊用途ハードウエアの1つの態様がアルゴリズムの実行を並列化することによる潜在的な利点を引き出すことを意図している。1つのノード（以下に定義する）がM（データの行の数）に沿った任意のデータセットを

分割し、これらの部分をそのCP（これも以下に定義する）に対して分配する。CPは他のノード（回帰的定義における）でもよく、上記の好ましい態様の節のプログラム法の説明の項で高レベル「疑似コード」において説明した方法のステップ8を実施するために開発された特殊用途プロセッサでもよい。ノードのCPによって結果が計算された場合、マージのステップ（上記の「疑似コード」の記載におけるステップ9から14まで）がノードによって実施される。いったんマージがなされた時点で、結果はノードのペアレント（parent）に戻される。ノードがツリーのルートでない場合には、このハードウエアを制御するドライバに完全な結果のセットが返される。以下に説明するシステムは主コンピュータのCPUから「オフライン」で用いうる。このようなシステムの商業的販売および使用に関するその他の可能性の中には、ユーザーが購入して自らのパーソナルコンピュータまたはワークステーションにインストールしうる特殊な「ボード」または「カード」上への実装がある。ローカルエリアネットワークまたは「スーパーコンピュータ」設備上での1つまたは多数のこのような特殊なサブシステムの使用も想起しうる。説明した態様は、当業者には理解されるであろうが、本明細書に記載された方法を並列化するための多くの可能な方式のうち1つのみを表現している。

以下に説明するこの履行（implementation）は、文字を値とするデータ属性の

みに対して作用すると仮定される。これは決して本明細書に記載される基本的方法を限定するものではなく、むしろこれは基本的方法の特殊な履行である。この実施は、本明細書の別項で記載される二値属性コード化に容易に追従しうる。

ノードの図は、図14に演算処理プロセッサ（CP）とともに示されている。ノードには以下のものが含まれる：

CPに送ろうとする入力が保存される（入力バッファー）、およびCPによって見いだされた結果が保存される（出力バッファー）記憶装置。

バス自体上の通信の調停に用いられるほかデータ転送の手段でもある、制御、データおよびアドレスバスに分割されたメモリバス。

一組のビットフラグおよび小さな追加メモリ部分（Lastout）。LastOutは最後に書き込まれた出力バッファーの区域のアドレスである。それぞれがどの状態にあるかを決定するためにマージおよびI／Oプロセッサによって2つのビットフラグ

が用いられる。

それぞれがそれ自体にローカルメモリキャッシュを備えた、一致の発見を実行する大きさJの演算処理プロセッサ（CP）のアレイ。

CPのマージ結果を書き込むメモリキャッシュをそれ自体に備えたマージプロセッサ（MG）。

バスの使用を制御することに主な役割がある入力／出力プロセッサ（IO）

システムの各要素があらゆる他の要素に関して同期的に実行されることを保証するために用いられるクロック。システムの各部分の実行は固定式に実行されるものと考えうる。

コンピュータプロセッサは、アルゴリズムのR-標本抽出ステップ（疑似コードにおけるステップ8および図5に図示されたもの）を実行する何らかの特殊なプロセッサと定義される。これは態様を単にベクトル配置に限定するのではなく、このようなノードのツリー構造の可能性をもたらす。メモリバス用ハードウエアの任意の特定の選択に関して、ノード当たりのCPの数に対して最大限に有用な制限があるという場合も考えられる。ツリー構造によりこの制限の回避が可能となる

。

この実施では、方法のパラメーターRおよびNの最大値（RmaxおよびNmax）が先験的に特定されると仮定している。これらの制限に違反が起こった際に検出し、それに応じて対応することはソフトウエアドライバの役割である。

２．記憶装置

各ノードに関してメモリのサイズは2*J*Amax*Rmax*Nmaxであり、ここでAmaxはノードにおいて行いうる反復の最大合計数である。このメモリは入力および出力バッファーに等しく分割される。単位の反復に関する入力の大きさはJ*Rmax*Nmaxを超えず、局所的に生成された結果および最終的にマージされた結果（J個のCPによる部分的結果をマージすることによって形成される）はいずれもこの制限を超えることができず、このため使用可能なメモリを上回るリスクがないことに注意されたい。

このメモリに対するアクセスは以下の通りである：

IOは入力バッファーに対する書き込みアクセス、および出力バッファーに対する読み取りアクセスを有する。

MGは入力バッファーに対するアクセスを持たず、出力バッファーに対する読み取りアクセスを有する。

CPは入力バッファーに対する読み取りアクセス、および出力バッファーに対する書き込みアクセスを有する。

３．メモリバス

メモリバスの制御はIOプロセッサの役割である。各CPには数値識別子が割り当てられる（0からJ+1ではIOに絶対的に0が割り当てられ、MGには1が割り当てられる）。メモリバスは3つの区域に分けられる。

制御：各CPに対する2本の線（wire）、MGに対する2本およびIPに対する2本が制御バスを構成する。各対の1番目は要求線（request wire）と呼ばれ、2番目は応答線（response wire）として知られる。

アドレス：システムの各装置には一意のメモリアドレス範囲が割り当てられる。アドレスバスはデータバスと組み合わせて用いられ、データバス上の現在の値

をどの装置に書き込むか、および適用可能であればその装置内部のどこに保存されるかを決定する。アドレスバスの幅（すなわち、内部の線の数）は入力および出力の記憶保存のために選択された大きさに関して決定され、このためここでは特定しない。

データ：文字を値とする属性のみがこのシステムによって処理されると仮定すると、データバスの幅は線8本分になる。

バスの調停（arbitration）は制御バスの使用を介して処理される。デバイス（ここではMG、IOまたはCPの1つを意味する）がバスを用いようとすると、それはその要求線に論理値1をアサートする。任意のサイクルに対して、複数のデバイスがそれを行いうる。IOはバスに調停動作を戻す際に、最も少ない番号のデバイスの応答線を1に、他のすべての応答線を0に設定する。これは最も少ないと同定されたデバイスに対し、それがバスを用いる許可を得ており（読み書きは指示されない—IOはこの状況を確立する役割を果たす）、それ以外のすべては待機する必要があることを伝える。バスを用いようとするすべてのデバイスは、許可を与えられるまで要求線に1をアサートしつづける。許可されたデバイスがバス使用を終了

すると、デバイスは要求線に0をアサートし、IOに対してバスを別のデバイスに対して再び割り当ててよいことを指示する。「ハンドシェイク」および上記のものなどの他の種類のプロトコルは当業者に周知であり、理解されている。

4．ビットフラグおよび追加メモリ

追加メモリは、IOにより、最後に書き込まれた出力区域を保存するために用いられる。出力バッファーに対する「書き込み」は段階的になされ、MGはその最後の読み取りインデックスを最後の書き込みインデックスと比較することにより、待機している未使用区域の数を決定しうるため、MGについてはこのような区域のリストを保存する必要はない。このメモリに書き込めるのはIOのみであり、それを読めるのはMGのみである。

「IOの終了」（IOがすべてのデータを送り出し、すべてのCP出力を受け取ったことを意味する）および「マージの終了」を示すために、2つのビットフラグが

用いられる。

５．J個の演算処理プロセッサのアレイ

上記の通り、本発明の一般的方法のアルゴリズム記載における1つのR-標本抽出ステップを計算するノードまたは特殊用途プロセッサが存在する。後者の場合、それは以下を含みうる：

以下に記載する関数に加えて一致検出を行うプロセッサ

サイズ2*Nmax*Rmaxのローカルメモリ

メモリは入力および出力に関する2つの等価な部分に分割される。

まず、CPがその要求線に1をアサートし、データに対する準備が整ったことを指示する。それが以下のサイクルの1つに対して設定された1つの応答線のみを監視している場合には、それはRおよびNに関する現在の値を送り、続いてデータ自体を送ると考えられる（さもなければ、これがそうなるのを待つ）。最初の2つの値に基づき、それは現在の入力が終了した時点を判断しうる。それは続いて要求線に0をアサートし、本方法の二値列化（binning）および一致検出のステップを行う。これらのステップが完了し、CPが論理値1を再び要求線にアサートした場合には、今回はそれがその結果を送ろうとしていることを示す。バス使用の許可が与えられると、それは一致集合をIOに送る。IOにはこのデータの保存場所を管理す

る役割がある。CPの出力ストリームは、一致（cset）自体によって見いだされた一致の符号を含む。一致は以下の形式をとる：

ヒット数（Rmaxを超えない）

サイズ（csetの幅、すなわち構成要素となる属性の数）

（値、位置）形式での一致の属性のサイズの長さのリスト

すべてのデータがIOに送られると、CPはその線に対してさらにデータを送るようにアサートする。

６．マージプロセッサMC

マージプロセッサは以下を含みうる：

マージのステップを実行するプロセッサ

1つのCPからの出力を保存するために用いられるNmaxRmaxのローカルメモリ

カウンタC1およびC2（前者はMGによって読み取られた最後の出力区域を探知し、後者はマージバッファーに現在保存されている一致の数を計数する）

Aの現在の値を保存するために用いられるメモリ

マージ結果を保存するために用いられるサイズJNmaxRmaxAmaxのメモリ

まず、MGはそのカウンタを0に、その要求線を0に設定し、IOが処理すべき出力データがあることをそれに対して知らせるまで（この線を1に設定することにより）待機する。

MGがその要求線がonになったことを認識すると、それはカウンタによって索引が付けられた出力データをそのローカルメモリに受け取り始める。いったんこれが達成されると、MGがマージアルゴリズムを開始することができる。マージはローカルメモリから直接マージバッファーに送られる形でなされる（このステップが終了した時点でC2は現在の一致の数を保持している必要がある）。このステップが完了すると、MGはLastOutの現在の値を検索して取り出す。それがC1よりも大きい場合、MGはそれがC1を増分として、次の出力区域に直接移動しうることを知る。C1とLaStOutが等しい場合、MGは要求線を0に設定する。C1がA*Jに到達した場合、MGはすべての結果が計算およびマージされたこと（およびこのためにすべてのCPおよびIOが休止状態にあること）を知り、このノードのペアレントを伝達するためにマージバッファーの内容をIOに送り戻す。結果は単にC2の値として送ら

れ、続いてマージバッファーに一致のリストが保存される（一致の形式は上記の5節で説明したものと同一である）。

7．入力／出力プロセッサIO

IOは以下を含む：

サイズJのビットベクトル

次に入手可能な出力二値列を示すカウンタC1

次の未使用の入力のR*N部分を示すカウンタC2

IOには上記に概要を述べたバス調停方式のための役割があり、アルゴリズムの

全体的な実行を司る。まず、IOはC1およびC2を0に設定し、そのビットベクトルを0にして（それがどのCPにもデータを送っていないことを示す）、ソフトウェアドライバがそれにデータを送るのを待つ。この間にそれは何ら動作をなし得ないことを知り、バスに対するすべての許可をゼロにする。割り込みがドライバからのデータの到着を知らせ、すべてのデータが入力バッファーに書き込まれるまで、IOはすべての通信要求をゼロにしつづける。入力データは以下の形式をとる：

N

R

T、送られたサイズRの行のセットの総数

サイズTRNのデータストリーム

IOはこのため、より多くのデータを期待することができない時点を決定する。以下を行うことはドライバの役割であることに注意されたい：

データマイニング要求をAmaxを超えないサイズに分割する

入力として送られた行の数がRによって割り切れることの保証

現在のデータセットがRmaxおよびNmaxを上回らないことの保証

デバイスから送り戻されたすべての結果のマージ

いったんすべての入力が保存されると、IOは、まずベクトル中のi番目のビットを1に設定し（これはIOがCPiからの出力を予想すべきであることを示す）、その応答線を1に設定してその他のすべてを0にすることによってCPに信号を送り、バスにデータを送り、最後にC2を増分とすることにより、サイズR*Nのデータを各CPiに送る。

すべてのCPがビジー状態である場合（またはすべての使用可能な入力が完了した場合）、IOはCPがその要求線に対して1をアサートするのを待つが、これはそれが結果を送り戻す準備ができたことを示す。いったんこの信号をCPから受け取ると、IOはCPから結果を検索して取り出し、カウンタにより索引が付けられた出力区域にそれらを保存し、そのCPに関連するビットを0にし、C1を増分として加え、MGの要求線に1をアサートする。入力バッファーに未使用データがある場合

、IOは次に使用可能なR*Nのセットを、結果を戻したばかりのCPに送る（そのCPに対するビットを1に設定する）。C2がTと等しく、ビットベクトルが1に設定されたビットを含まない場合、IOはそれが終了したことを知り、IOビットフラグを1に設定する。この時点で、MGビットも1に設定されたこと（MGがその動作を終了したことを示す）を見るまで、IOは前記の待機状態に復帰する。いったんこれが起こると、IOは割り込み（このノードがツリーのルートである場合）または単なる送付（このノードにペアレントの別のノードがある場合）の要求を呼び出し、MGにバスへの書き込みを許可し、続いてMGから送られたすべてのデータをペアレントに送り出す。

提案された方式はCP間で等しくない実行時間を可能とすることに注意されたい―データを得る次のCPは最終的なデータ許可を最後に得たものである。したがって、システムの全体的な動作はクロックによって制御されるものの、ある程度の非同期的処理能力もある。

特定のプロセッサ、バスおよびその他の構成要素に関する選択は、設計者、製作者、製造者、販売者、バイヤーおよびユーザーの裁量にまかされ、選択肢の範囲は当業者に周知である。特に、上記の態様のすべての部分は「標準仕様品」の材料から入手してもよく、または当業者によりVLSIレベルで特別に設計されたものでもよい。

種々の応用

概論

特殊用途の態様も可能である。例えば、マーケティングおよび販売／取引データの分析への適用において、本発明の方法に対する対象入力は取引に対応することができ、属性は特定の製品またはサービスの販売事例に対応する。

工程管理、生産工学またはコンピュータシステム管理への適用では、対象は特定の時間刻みまたは期間に対応することが可能であり、属性は特定の構成要素、リソースまたはサブシステムのオン／オフまたは使用／非使用状態に対応する。この適用の目標は、効率の改善または事業コストの削減のために、相互作用するサブシステムまたはユーザー間でのk項対立または対立する需要を見いだすこと

であってもよい。

例えば、本方法は、図8の全体的流れ図および図9の模式図に示したような製品の製造工程の制御のために適合化しうる。この例は自動化板金組み立て工場を表しうる。本方法を、プラントからの製品の1つに対する需要が周期的変動のために夏期には著しく減少すると思われ、一方で別の製品に対する需要は増加するという相関を見いだすために、既存のデータセットに対して適用することも可能であろう。プラントの自動化工程制御システムへの接続で最初の製品に対する発注を減らし、もう1つに対する発注を増やすことができる。相関が見いだされた結果としての製品の実際の構造へのバリエーションを含む、その他の多くの例が当業者には明らかであると思われる。

1つの代替的な態様では、規則に基づくシステムのための規則を作り出し、続いてそれらの規則に基づく製品を製造するために、見いだされた相関を用いうる。このような態様に関する全体的流れ図は図10に示されている。対応する模式図は図11に示されている。

さらなる代替的な1つの態様では、製品を製造する工程を制御するために、規則に基づくシステムを用いうる。このような態様に関する全体的流れ図は図12に示されている。対応する模式図は図13に示されている。

財政分析または取引への応用では、対象は特定の時間刻みまたは期間に対応することができ、変数は特定の金融証書または商品の特定の価格または価格変動に関するものでありうる。各証書または商品の価格を一組の離散的レベルに区分することにより、または「上昇対減少」に関する単純な二進符号を用いることにより、このようなそれぞれの証書または商品を属性の集合によって表現することができ、価格変動が相関するk項数の証書または商品を見いだすために本発明を用いることができる。当業者は、この種の見いだされた情報から価値を得るための多

くの方式を周知している。

医学、疫学または環境科学への応用では、対象は特定の患者、または一人の患者の種々の時期の観察結果、または同一もしくは異なる環境資源（空気、土壌ま

たは水など）からの試料に対応することができ、変数および導出された属性は特定の症状、薬物、毒素または汚染物質のレベルまたは存在／非存在に対応しうると考えられる。この方式では、本発明を用いて、疾患または環境公害の原因となりうる相互作用を見いだすことができる。

分子および構造生物学への応用では、対象はDNA、RNAまたは蛋白質の配列および／または構造に対応しうる。属性は、特定の配列位置での特定の塩基もしくはアミノ酸の存在、または特定の配列もしくは構造上の位置での特定の幾何学的、化学的、物理的もしくは生物的な性質、またはその他の全体的または局所的性質の有無もしくはレベルでありうる。例えば、以下に示すものは、蛋白質構造の予測のための本方法の詳細な応用であり、これはこれまでに説明したものの例である。

薬理学的な応用では、対象は特定の化合物または薬物の分子構造またはその他の標識または表現に対応することができ、属性は、例えば特定の化学成分などの、特定の幾何学的、化学的、物理的、生物的、毒性学的、治療的および／またはその他の性質および特徴の有無または程度に対応しうる。本方法は、k項数のこのような性質の間の相関を見いだすために用いうると考えられ、この情報は化合物および薬物の設計および試験、ならびにスクリーニングおよび試験のためのコンビナトリアルライブラリーの設計、または薬物探索および薬物設計におけるその他の工程または段階のために有用な可能性がある。または、上記の配置を転置して、対象が性質および特徴に対応し、属性が化合物および薬物に対応するようにすることも可能である。この方式では、類似的または相補的または相乗的または拮抗的活性を有する一連の薬物を見いだすために本発明を用いることができる。これもまた、薬物探索および薬物設計において極めて有用である。

人口統計学、マーケティング、保険および信用度格付け、ならびに／または資金調達への応用では、対象を特定の人々または会社または組織に対応させることができる。属性は、雇用、収入、財産、信用度の履歴、生活様式、消費パターン

または社会的／政治的意見もしくは所属に関する性質および特徴の有無または程度に対応させうる。本方法はこのような因子間の関連を見いだすために用いうる

と考えられ、それは信用度／保険リスクの予測もしくは詐欺の発見などの業務において、または例えば限定的なマーケティングの配分もしくは資産調達に最適なターゲットの判断において有用でありうる。

データベース中の対またはk項数の属性間のすべての有意な相関を見いだすという問題は、数理科学ならびに医学、工業および財政分野において普遍的である。本明細書に記載された原理は、N-属性データベースにおいて$2 \leqq k \leqq N$であるすべてのkに関して、正当に有意な対相関を見いだすのと同じ計算コストで有意な高次k項相関を見いだすという興味深い性質を有する確率論的アルゴリズムを含む。さらに、他の既知の手順とは対照的に、kは本発明者らの手順においてあらかじめ固定されている必要はない。本手順は整列化された蛋白質配列における保存された構造的関係を見いだす作業のために設計されたが、その他の分野でもより有用な用途がある可能性がある。

本明細書に記載された原理の蛋白質配列解析への応用

蛋白質鎖には配列が隔たったアミノ酸残基の間の相互作用がみられ、時に蛋白質構造ファミリーからの一組の整列化された配列における位置（列）の間の相関として検出可能であるが、これは構造および機能の決定に重要な役割を果たす。見いだされた相関は代償性変異の進化の歴史を表すことがあり、蛋白質の構造／機能ファミリーのモデルに有用な特徴を提供する可能性があるが、大部分のML（機械学習による）分類法では無視または誤った取り扱いがなされており、これは一部にはk項数の相関する位置を検索するという計算の高度の複雑性による。

ヌクレオチドまたはアミノ酸配列などの生物的配列の行列に対して本発明を実施するためには、選択的には比較の目的で異なる配列をまず整列化する。第1の配列における位置を第2の配列における対応する位置と比較する。比較した位置が同じヌクレオチドまたはアミノ酸で占められている時には、場合によっては、2つの配列はその位置で同一である。2つの配列間の一致の度合いはしばしば、2つの配列における合致（同一な）位置の数と比較した位置の総数との比を表す比率として表現される。選択的には、2つまたはそれ以上の配列の整列化には一般に、それ

らの間の配列一致の度合いの最大化が含まれる。

配列整列化の分野の当業者には、いくつかのアルゴリズムおよびコンピュータプログラムが知られている。これらのツールには、フェング（Feng）およびドゥーリトル（Doolittle）[J. Mol. Evol. 25, 351 (1987)]の漸進的整列化法の修正版を用いるGenetics Computer Group (Madison, WI) パッケージ（バージョン8）からのPILEUPプログラム、欧州分子生物学研究所（European Molecular Biology Laboratory）（EMBL）、Heildelberg、Germanyから入手可能なフリーウエアであるCLUSTAL X、および米国立衛生研究所（National Institutes of Health）（NIH）、Bethesda, MIDから入手可能なフリーウエアであるBLASTが含まれ、BLAST-Pはアミノ酸配列に対して用いられ、BLAST-Nはヌクレオチド配列に対して、BLASTXは核酸コドン／アミノ酸翻訳に対して用いられる。

蛋白質配列ファミリーの解析からはいくつかの種類の有用な情報を得ることができる。

第1に、結合記号の頻度の形で、個々の配列のレベルで抽出される情報がある。特定の単一位置パターン（例えば「これらの配列の98%で第3残基にGが生じる」）が異常に高い頻度で観察されることにより、二次または三次構造に対する重要な物理化学的拘束が判明する可能性があることが知られている。驚くほど高い頻度での連結記号の出現にも同じことが成り立つ（例えば「3位でのG、5位でのLおよび87位でのMが、個々の周辺頻度による予測よりもはるかに高い頻度で出現する」）。このような長い距離を隔てた同時出現は特に三次拘束を示すが、これは指定された位置が、モデル化された配列のすべてと対応する3D構造において互いに近接する可能性があるためである(p(A, B)>>p(A)p(B)である場合のこの「疑わしい一致」、の検出は、かなり以前に他の者によって指摘された通り、パターン認識および学習の核心である）。

第2に、位置（相同配列の整列化における列）の間の統計的関係に関して、「次に高いレベル」で抽出される情報がある。連結記号のk項数が出現する頻度の存在を3D構造相互作用の推論に用いることができる場合、多くの異なる連結記号の出現の集合にわたっての位置（列）の間の特定の情報理論的関係により、このような推論ははるかに強く裏づけら得れる。このため、このような記号的関係は

蛋白

質鎖の異なる部分の間での進化的に保存された物理的または構造的関係を示しうる（図15参照）。列の間の相互情報および他の相関の程度に高値が観測されることは、RNAおよびHIV蛋白質における3D構造相互作用の予測に好首尾に用いられており、例えばシャノン（C.E. Shannon）およびウィーバー（W. Weaver）、通信の数学的理論（The Mathematical Theory of Communica Lion）、The University of Illinois Press, 1964を参照されたい。これらの以前に発表された取り組みでは、対の残基-残基間相互作用に対象を絞っているが、本明細書で記載される原理は2≦k≦Nのk項相互作用の検出を目的とする。

発見されたk項数の相関するアミノ酸残基は、蛋白質の構造予測および構造決定に用いうる。

局所予測は、最良の全体的構造予測に対する検索の幅を狭めるのに有用と思われる。

第1に、距離幾何学的な拘束がある。二次構造の予測およびk項長距離相互作用の発見は、蛋白質中のi番目およびj番目のアミノ酸残基に関してcontact(i, j)の形式をとる、推定された接触に対する証拠となる。この種の距離幾何学理論は他の者によって開発されており（例えば、T.F. Havel, L.D. Kuntz, G.M. Crippen、距離幾何学の理論および実践（The Theory and Practice of Distance Geometry）、Bull. of Mathematics Biology v.45 1983pp.665～720およびK.A. Dill, K.M. Feibig, H.S. Chan、蛋白質フォールディング動態における協同性（Cooperativity in Protein-Folding Kinetics）、Proc. Natl. Acad. Sci. U.S.A. v.90 March 1993 pp.1942～1946を参照されたい）、推論された接触の集合を誘導することができる。推定または推論された接触の任意の集合によって禁じられる接触である、推論されたブロックの集合を誘導することも可能である。本質的には、固定体積内に存在するように拘束された重合体鎖のモデルが与えられたとして、2つの特定の小片が接触に至ることは、いくつかの他の小片は近接するようになるが他の小片はさらに離れることを意味する。事実、別の者は「密集した重合体では、単に立体的制限により、かなりの量の内部構造（ヘリックスならびに

平行および逆平行シート）が生じることが予測される。これは球状蛋白質にこれだけ多くの内部組織化がみられる一因と考えられる」と結論している。

第2に、以前の節の全体を通じて考察した通り、局所的および全体的な立体配置の間の経験的な関係を推論または利用することができる。配列の局所的範囲、または選択された非局所的残基対は、いくらか高い確率で、特定の球状立体配置において生じることを見いだしうる。帰納的規則はいかなる形式のものでも、立体配座空間の大きな部分を避けるために用いうる。特定のフォールディングにおける協同性モデルの推論は特殊な例である：p(contact(i,j)|contact(i+1,j-1))>p(contact(i,j))などの「規則」の知識は極めて有用なことがある。

例えば、図16は三次構造予測における段階を図示している。本出願の全体を通じて記載される方法は、比較的大きな三次構造予測システムの一部として適用することができ、ここで上記の原理は整列化された配列ファミリーの解析に関するブロックにおいて用いられる。このシステムは蛋白質の構造を予測する。

進化的に保存された構造的拘束の発見

この節では3つの問題を取り扱う。

1. 多数の配列アライメントにおける列の間の相関を検出することにより、いかなる種類の進化的に保存された多残基構造的または機能的拘束が見いだされると考えられるか？

2. 相関-検出の取り組みは実際に重要な構造的または機能的拘束を見いだしているか？

3. このような発見は、分子の本来の三次構造の予測または決定に向けてどの程度の情報を提供するか？

我々が観察しうると考えられるものは何か？

蛋白質ファミリーとは、共通の全体的三次構造を共有すると考えられる一組のアミノ酸配列である。蛋白質のフォールディングおよび進化の理論および観察により、蛋白質ファミリー内部での進化および保存という一般的な概念が裏づけられている。

・機能的拘束は表面残基に保存される、

・構造的拘束はコア残基に保存される、

・変異性浮動は主にループ残基にみられる

機能的拘束には、他の蛋白質、核酸、脂質、金属、O_2または他の低分子などの他

の分子がしばしば関与する。

蛋白質ファミリーの進化を通じて保存されると考えられる種類の構造的拘束は、立体配座を安定化する2、3の重要な残基が主として関与するものである。静電相互作用が重要と思われる場合には、2つまたはそれ以上の配列位置にわたり実効電荷の保存が見いだされると予想される。静電的に相互作用する2つの残基のうち1つが正電荷を有する場合には、その「パートナー」残基（配列では離れていても3D構造ではおそらく近接する）は負に荷電している必要があり、その逆もしかりである。状況はパッキング拘束についても類似している。蛋白質コア体積の部分は、同じ構造ファミリーに属する多くの異なる蛋白質を通じてわずかな差しかないが、非コア領域は大きな体積変動を呈すると正当に考えられる。したがって、側鎖体積に関して互いに代償的な変異を呈する対または少数k項数の残基が見いだされる、すなわち簡単にいえば「大」が「小」に変異した場合には別の「小」が「大」に変異する必要があると考えることができる。

何が観測されているか？

ネーアー（Neher）ら（蛋白質配列ファミリーにおける相関的変化はどの程度の頻度か（How frequent are correlated changes in families of protein sequences）、PNAS, 91：98～102, 1994）は、アミノ酸に関する物理化学的性質の指標を用い、続いてアライメントにおける列間のピアソン相関を評価することにより、単一の蛋白質ファミリーの内部での代償的変化の頻度を定量化しようと試みた。彼らは、ファミリー由来の対の配列の検討に基づき、ブートストラップを参考にした再標本抽出方式を用いて小規模データセット問題を回避しようと試みた。ミオグロビンファミリーの蛋白質配列に対する彼らの研究により、側鎖体積の性質に対する代償的変異の程度は低いが、電荷に対しては程度が高く、局所電荷が完全に保存されるとして予想される相関レベルに近いことが明らかになった

。著者らは、彼らの列一対解析は接触性の近傍残基対のみに対象を絞っているため、電荷保存のように極めて局所的に作用する拘束を見いだすことはできたが、体積の保存のようにより分布の広い拘束は検出しえなかったと推測している（換言すれば、正に荷電した単一の残基は負に荷電した単一の構造的パートナーと接触する必要があるが、適合性の体積をもつ一組のパートナーは複数の残基を含む可能性

があり、すべてが接触する必要はないと思われる）。別の者も、蛋白質構造ファミリーの進化における協同的変異の若干の証拠を見いだしている。

代償的変異に関する今日までの大部分の研究は、蛋白質構造の高度に保存された「コア」型領域に対象を絞っているが、ケルバー（Korber）ら（HIV-1のV3ループにおける変異の共変：情報理論による分析（Covariation of mutations in the V3 loop of HiIV-1:An information-theoretic analysis）、Proc. Nat. Acad. Sci, 90, 1993）は、HIVエンベロープ蛋白質の高度に可変的なV3ループを分析した。この研究者らは、V3の残基を表す31列の組からのすべての列の対に関して対変異情報の強力なブートストラップによる推定を行った。彼らは、かなりの統計的に有意な変異情報を示す約7対の組を見いだし、特定の属性（アミノ酸）の解析により、極めて可能性の高い代償的変異の特定パターンが示唆された。著者らは保存された特定の性質または関係に関する議論もその証拠の提供も行っていないが、研究室でのその後の変異解析では高度な変異情報を有する部位の対の一部の間に機能的連鎖が認められた。V3領域は機能的にも免疫学的にも重要であることが知られており、本出願の発明者らはこの種の解析がHIV／AIDSワクチンのデザインの検索に重要であろうと示唆している。

<u>いかなる種類の方法が必要か？</u>

明らかに、蛋白質配列ファミリーの包括的モデル化のためにはいくつかの詳細に研究された有効な方法が存在する。それぞれの場合において、数学的手続きはデータにおける極めて局所的および低次統計的構造の取り扱いおよび検出に適する。それぞれの場合において、残基間、すなわち整列化された配列データにおける列の間のすべての可能な非局所的および高次相互作用を総体的に追及しようと

いう試みは、計算上の複雑性および統計的推定に伴う困難さが生じる。

HMMまたは密度ネットワークを、信頼性のある非局所的相互作用の検出に明確に対象を絞り、これらの相互作用のモデル化における精度の程度を犠牲にする迅速な帰納的プレプロセッサとともに用いれば、より容易にモデル化を進めることが可能である。このような手順は本明細書に記載される原理によって提供される。

a) HIVタンパク質の配列分析

HIVタンパク質データベースの検定

ロス・アラモスHIVデータベースは、その中にHIVエンベロープタンパク質のV3ループ領域のアミノ酸配列をも含んでいる。この領域は、機能的および免疫学的な重要性をもつことが知られており、また、進化の共変につながる部位集合の発見はHIVの感染および複製を理解および防止する上での重要な示唆を含む可能性がある。

同一のデータベースの、初期の小規模版を用いて、ロス・アラモスの科学者らは残基間の対相互情報量の解析をおこなった。

HIVデータ集合上で、一致検出法を用い、rおよびTの様々な数値の集合に関して実験を行った。結果を示す表およびそれらの考察を以下に示す。

HIVタンパク質データベースの実験結果

ロス・アラモスの研究員は、最も保存性が高く、また機能的および免疫学的に重要であると考えられる３３残基について集中的に調べるために、先述のHIV-V3データ集合版を編集した。従って、データ集合はM＝６５７行（配列）およびN＝33列（残基）から成る。一致検出法では、33列はNA=N. ＼A＼=33.21=693属性に変換した。人工のデータ集合と同様にして、様々なTおよびrの値を用いて一連の実験を行った。一致検出は先ず、T=10,000かつr=5、6、7、10でそれぞれ行い、その後、T=100,000かつr=7、最終的にはT=750,000かつr=7で行った。結果を以下の表C.1からC.9に示す。

表C.1　一致検出法により推定された、HIVデータ集合における最も相関の高い可

能性のある属性。これらの結果はT=10,000、r=5のパラメータで求められた。

HIVデータ集合

T=10,000、r=5

順位	ＣＳＥＴ	実測値	予測値	確率
1	*Q*17|*D*24	1012	632.553864	0.316056
2	*R*17|*T*21	901	610.770465	0.509734
3	*R*12|*Q*17	570	348.605833	0.675621
4	*L*13|*W*19|*Q*24	195	5.535741	0.750381
5	*N*4|*K*9|*A*21	226	74.167398	0.831582
6	*V*11|*R*12|*T*18	159	20.764346	0.858239
7	*R*12|*T*18	454	318.517747	0.863429
8	*L*13|*K*31	419	300.333903	0.893461

表C.2　一致検出法により推定された、HIVデータ集合における最も相関の高い可能性のある属性。これらの結果はT=10,000、r=6のパラメータで求められた。

HIVデータ集合

T=10,000、r=6

順位	ＣＳＥＴ	実測値	予測値	確率
1	*Q*17|*D*24	1177	385.853329	0.030891
2	*R*17|*T*21	957	368.736702	0.146238
3	*H*12|*A*18	1047	577.583832	0.294000
4	*S*10|*D*24	859	424.457490	0.350274
5	*R*12|*Q*17	656	224.743830	0.355855
6	*R*12|*T*18	628	283.191527	0.516585
7	*R*17|*E*24	563	234.477161	0.549033
8	*H*12|*R*17	760	434.274580	0.554644
9	*A*18|*T*21	560	315.973734	0.718330
10	*I*11|*R*17	861	627.014684	0.737741
11	*L*13|*W*19|*Q*24	230	5.365202	0.755529
12	*A*21|*D*24	619	405.487239	0.776262
13	*N*4|*K*9|*A*21	237	25.176801	0.779367
14	*V*11|*R*12|*T*18	220	15.841474	0.793296
15	*L*13|*K*31	462	267.211446	0.809942
16	*G*10|*H*12	324	157.554658	0.857348
17	*M*13|*W*15	245	84.760597	0.867059
18	*Q*17|*K*31	384	231.749746	0.879169
19	*H*12|*R*17|*A*18	147	8.219536	0.898526
20	*N*4|*K*9|*H*33	309	170.353419	0.898711

表C.3　一致検出法により推定された、HIVデータ集合における最も相関の高い可能性のある属性。これらの結果はT=10,000、r=7のパラメータで求められた。

HIVデータ集合

T=10,000、r=7

順位	ＣＳＥＴ	実測値	予測値	確率
1	*Q17|D24*	1312	228.829775	0.008322
2	*N4|K9*	2023	996.505631	0.013558
3	*H12|A18*	1175	328.263693	0.053591
4	*R17|T21*	940	216.431391	0.118015
5	*Q31|H33*	3198	2481.050915	0.122699
6	*R12|T18*	879	244.789294	0.193645
7	*S10|D24*	836	232.201517	0.225812
8	*R12|Q17*	720	140.866087	0.254370
9	*I11|R17*	808	360.719364	0.441944
10	*H12|R17*	659	253.717115	0.511491
11	*R17|A118*	720	361.819054	0.592356
12	*A21|D24*	554	236.085429	0.661974
13	*R17|E24*	452	138.843412	0.670137
14	*L13|K31*	537	231.137972	0.682602
15	*L13|W19|Q24*	292	5.055474	0.714573
16	*A18|T21*	442	165.231990	0.731502
17	*A18|Q31|H33*	480	209.122778	0.741198
18	*M13|W15*	355	88.975694	0.749122
19	*N4|K9|H33*	340	75.556215	0.751690
20	*V11|R12*	513	253.001684	0.758878

表C.4　一致検出法により推定された、HIVデータ集合における最も相関の高い可能性のある属性。これらの結果はT=10,000、r=10パラメータで求められた。

HIVデータ集合

T=10,000、r=10

順位	ＣＳＥＴ	実測値	予測値	確率
1	*Q31|H33*	3933	883.532458	0.000000
2	*N4|K9*	2898	251.248235	0.000001
3	*S10|F19*	2245	907.769718	0.027977
4	*F19|G23*	2660	1588.173503	0.100497
5	*R12|T18*	1155	142.229768	0.128554
6	*K9|I11*	1230	311.653160	0.185125
7	*A18|H33*	1720	990.576490	0.345032
8	*K9|H33*	1125	405.874883	0.355482
9	*H12|A18*	732	54.213558	0.399002
10	*S10|G23*	1492	856.152048	0.445479
11	*N4|H33*	1257	689.784961	0.525468
12	*A18|Q31*	1188	636.901303	0.544755
13	*Q17|D24*	571	42.938312	0.572525
14	*V11|R12*	670	143.659674	0.574607
15	*I11|R17*	562	61.788305	0.606274
16	*N4|R17*	992	498.586806	0.614520
17	*R12|Q17*	484	31.204991	0.663619
18	*K31|Y33*	578	130.131866	0.669535
19	*R17|T21*	479	39.372545	0.679400
20	*S10|D24*	451	34.199456	0.706491

表C.5　一致検出法により推定された、HIVデータ集合における最も相関の高い可能性のある30属性。これらの結果はT=100,000、r=7のパラメータで求められた。

HIVデータ集合

T=100,000、r=7

順位	ＣＳＥＴ	実測値	予測値	確率
1	*H12|A18*	11686	3282.636926	0.000000
2	*N4|K9*	21853	9965.056308	0.000000
3	*Q17|D24*	11585	2288.297747	0.000000
4	*Q31|H33*	31715	24810.509148	0.000000
5	*R17|T21*	9355	2164.313906	0.000000
6	*R12|Q17*	7259	1408.660868	0.000001
7	*R12|T18*	8380	2447.892936	0.000001
8	*S10|D24*	7666	2322.015166	0.000009
9	*I11|R17*	8336	3607.193645	0.000109
10	*A21|D24*	6342	2360.854285	0.001550
11	*H12|R17*	6363	2537.171146	0.002543
12	*R17|A18*	7162	3618.190543	0.005941
13	*R17|E24*	4451	1388.434119	0.021747
14	*A18|T21*	4673	1652.319901	0.024130
15	*V11|R12*	5486	2530.016841	0.028256
16	*L13|K31*	5224	2311.379719	0.031348
17	*N4|K9|H33*	3519	755.562151	0.044291
18	*A18|Q31|H33*	4665	2091.227775	0.066951
19	*L13|W19|Q24*	2585	50.554739	0.072672
20	*R17|Q31*	5967	3574.032278	0.096592
21	*M13|W15*	3204	889.756945	0.112364
22	*V11|R12|T18*	2424	117.500168	0.114017
23	*N4|A21*	6209	4030.321314	0.144077
24	*K31|Y33*	4878	2773.817984	0.164117
25	*Q17|K31*	3440	1450.098718	0.198651
26	*K9|A21*	5614	3692.671816	0.221632
27	*P19|D24*	3998	2250.071839	0.287354
28	*Q17|A21*	4151	2414.536189	0.292077
29	*G10|H12*	2661	953.572593	0.304245
30	*H12|E24*	3018	1458.576938	0.370622

表C.6　一致検出法により推定された、HIVデータ集合における最も相関の高い可能性のある50項目のうち上位25位の属性。これらの結果はT=750,000、r=7のパラメータで求められた。この標本数では、k≧3で、いくつかの統計学上有意な高位の特徴が現われていることに着目。

HIVデータ集合

T=750,000、r=7

順位	ＣＳＥＴ	実測値	予測値	確率
0	*A18|Q31|H33*	36019	15684.208314	0.000000
1	*A18|T21*	33816	12392.399254	0.000000
2	*A21|D24*	45549	17706.407140	0.000000
3	*H12|A18*	86025	24619.776947	0.000000
4	*H12|R17*	48257	19028.783592	0.000000
5	*I11|R17*	64548	27053.952336	0.000000
6	*L13|K31*	39382	17335.347894	0.000000
7	*L13|W19|Q24*	20184	379.160544	0.000000
8	*M13|W15*	23300	6673.177086	0.000000
9	*N4|K9*	162152	74737.922307	0.000000
10	*N4|K9|H33*	26376	5666.716129	0.000000
11	*Q17|D24*	86891	17162.233105	0.000000
12	*Q31|H33*	233190	86078.318611	0.000000
13	*R12|Q17*	53740	10564.956512	0.000000
14	*R12|T18*	62774	18359.197022	0.000000
15	*R17|A18*	54366	27136.429076	0.000000
16	*R17|E24*	33748	10413.255892	0.000000
17	*R17|Q31*	45065	26805.242087	0.000000
18	*R17|T21*	70301	16232.354294	0.000000
19	*S10|D24*	57772	17415.113746	0.000000
20	*V11|R12*	39546	18975.126308	0.000000
21	*V11|R12|T18*	17628	881.251263	0.000000
22	*K31|Y33*	36346	20803.634880	0.000002
23	*N4|A21*	45441	30227.409858	0.000003
24	*Q17|K31*	25033	10875.740384	0.000018
25	*G10|H12*	20779	7151.794446	0.000041

表C.7　一致分析法により推定された、HIVデータ集合における最も相関の高い可能性のある26位から50位の属性。これらの結果はT=750,000、r=7のパラメータで求められた。この標本数では、k≧3で、統計学上有意な幾つかの高位の特徴が現われていることに着目。

HIVデータ集合

T=750,000、r=7

順位	ＣＳＥＴ	実測値	予測値	確率
26	*K*9|*A*21	40098	27695.038620	0.000231
27	*F*19|*D*24	29121	16875.538795	0.000286
28	*Q*17|*A*21	29621	18109.021417	0.000737
29	*H*12|*E*24	22348	10939.327036	0.000839
30	*N*4|*K*9|*V*11	15175	4159.316971	0.001355
31	*S*4|*T*9|*T*12|*V*18|*R*21	10919	1.718549	0.001524
32	*N*4|*K*9|*A*21	11233	623.181959	0.002185
33	*N*4|*Q*31|*H*33	21868	11328.342993	0.002369
34	*F*19|*A*21	44400	34516.144368	0.004910
35	*K*9|*Q*31|*H*33	16593	6991.723718	0.006625
36	*W*19|*Q*24	16738	7234.038664	0.007331
37	*E*1|*N*12	10844	1492.835945	0.008575
38	*K*9|*E*24	13847	4587.312260	0.009408
39	*K*9|*R*17	33735	24568.179150	0.010326
40	*T*12|*V*18	23076	14893.617567	0.026158
41	*R*12|*A*21	15497	7516.155896	0.031231
42	*N*4|*K*9|*Q*31|*H*33	8280	493.681367	0.036905
43	*N*4|*K*9|*A*18	11655	4250.900600	0.050618
44	*S*4|*T*9|*T*12|*V*18|*R*21|*Y*33	7370	0.093039	0.052029
45	*R*12|*Q*17|*T*18	7452	240.364918	0.058992
46	*V*11|*Q*17	14350	7329.962834	0.068429
47	*H*12|*T*21	23263	16324.923094	0.072825
48	*Q*17|*Y*33	17288	10374.788061	0.074203
49	*L*13|*W*19	15536	8921.243955	0.092437
50	*S*17|*H*28	6529	138.997153	0.108375

表C.8　本文中に記された方法により推定された、HIV-V3データ集合の対列間相互情報量の上位３５位までの値。

順位	対 i,j	MI(c_i,c_j)	標準誤差
1	12|18	0.340449	0.037792
2	4|9	0.337943	0.0389162
3	9|21	0.319481	0.0353829
4	23|24	0.315202	0.0337213
5	12|24	0.314393	0.0330382
6	9|24	0.313992	0.0344732
7	19|24	0.305609	0.0335857
8	11|24	0.297498	0.0358645
9	24|26	0.290044	0.0384839
10	9|11	0.289911	0.0344244
11	9|23	0.285019	0.0343224
12	4|21	0.284936	0.0332236
13	18|21	0.278151	0.0404634
14	4|11	0.277189	0.0353993
15	12|21	0.273137	0.033385
16	4|24	0.262226	0.036189
17	21|24	0.260366	0.0338395
18	11|23	0.260337	0.0323302
19	11|19	0.249877	0.0320634
20	10|24	0.248938	0.0325318
21	19|23	0.242185	0.032301
22	5|26	0.239395	0.0386373
23	9|19	0.238318	0.0331283
24	4|23	0.23359	0.0302795
25	24|25	0.222109	0.0358744
26	6|26	0.220371	0.0397722
27	4|26	0.220213	0.0333324
28	6|24	0.218815	0.0335123
29	9|12	0.214844	0.0280984
30	15|24	0.213921	0.0301834
31	10|12	0.2133	0.0306496
32	9|18	0.21078	0.031734
33	11|21	0.210155	0.0308121
34	11|12	0.209421	0.0294066
35	4|19	0.20911	0.0290533

表C.9　ロス・アラモスのグループの推定による、対列内相互情報量の上位７位

のHIV-V3データ集合の値。

順位	対 i, j
1	23|24
2	12|24
3	12|18
4	12|23
5	19|24
6	10|24
7	10|12

表C.1からC.4は、各属性の検出された一致について観察された一致数においての(本発明者らの方法による確率（実測値／独立事象）の推定によって測定された)最も有意なCSETを示す。予期されるように、この現実のデータ集合について、この比較的小さな標本抽出数では、「おそらく相関がある」と「おそらく相関がない」との間の明確な区別は見られない。r=7およびr=10での結果は、r=5およびr=6の結果よりも、検出されたCSETb5より有意であることを示している。前者の高いr値で、確率が0.1より小さい組み合わせは、(Q@17, D@24)、(N@4, K@9)、(H@12, A@18)、(Q@31, H@23)、および(S@10, F@19)である。これらのCSETは全て、注目すべき例外である(S@10, F@19)の組み合わせ以外は、より集中的な標本抽出(T=100,000およびT=750,000)に報告されている、最も有意なCSETに含まれている。この後者のCSETは、小さな標本抽出の度合いで、r=10でのみ見

つかっているが、r=7を用いた場合の、より集中的な標本抽出の種類では見られない。

表C.5は、T=100,000、r=7についての結果を示している。ここでは確率が0.1以下の有意水準内で、１７の対および3つの3項（3-ary）相関をともない、HOF集合内C、雑音と信号の区別がいくらかおこっていることが明らかである。

表C.6およびC.7中に示されるように、T=750,000において、より多くの統計学的に有意な組み合わせが検出され、５０近くの2項、3項から6項までの属性の相関が見られた。

これらの結果が意味するところを明確に把握するために、本発明者らの独自分析およびロス・アラモスグループによる対相互情報量を推定して、列間相関と共に、これらの属性間相関を考察することとする。表C.8は本発明者らの33列のデータ集合からのN-N=528のすべての組み合わせ中の、最も高い推定対情報量値を示している。これらの推測値は、M=657のうちm=300の1000の標本データの部分集合を抽出し、標準対情報量の計算法を用いるブーストラップ法（Bootstrap）のような方法により得られた。従って、表中では、標本についての平均値と、それに関連する標準誤差値が報告されている。表C.6およびC.7中の最大のCSET値によって示された列の組み合わせの集合と表C.8中の最大の対情報量値で示された集合との間に有意な交事象が存在する。二つの順位間の対応は、幾つかの理由（雑音および単純な標本の誤差以外の）により完全ではない。第一の、そして主要な理由は、単一の結合属性の組み合わせの「疑惑」は確実に、対応する列の集合内での対情報量に貢献する一方、列内に現れるその他の記号の行動もまた、明らかに大きな影響をもつことができる。次に、観察された感受性一致検出は、rの選択に帰すことを再度言及する。

表C.9にはロス・アラモスグループにより推定された、統計学的に最も有意な対情報量が記載されている。この表と本発明者らの表との間の共通部分に着目するものの、ロス・アラモスグループはより初期の小規模なデータベースを、おそらくさもなくば本発明者らがアクセスしなかったデータベースを使用したことを再度強調したい。

こうして、本発明の一致分析法を、このように整列されたHIV配列のような生物

学的データに応用することは、かつて認識されていなかった共変する構造要素を同定することにつながる。構造および機能は生化学系において密接に関連しているため、アミノ酸残基のような特定の構造要素の統計学的に有意な一致は、共変する構造要素を有するモチーフの生物学的役割を示す可能性が高い。本発明の上記の応用の一つには、HIVエンベロープタンパク質のV3ループにおけるA18、Q31、およびH33残基の統計学的に有意な一致がある。これらの残基は、HIVのライフ

サイクルにおいて生物学的役割を果すV3ループの構造モチーフに貢献することが期待される。本発明以前には、特定の生物学的役割としてまとめられることがなかったA18／Q31／H33についてのこのような新しい情報は、以下のように、様々な方法で開拓される可能性がある。

本発明により、ペプチドまたは、先に述べたV3ループ構造モチーフ（または一致検出法により同定された他のタンパク質モチーフ）に似た疑似ペプチドが提供された。選抜した例について、A18／Q31／H33アミノ酸の全ての原子は必ずしも必要ではないが、ペプチド、または疑似ペプチドにはこれらのアミノ酸残基の空間配位が含まれる可能性がある。V3と他の生物学的分子との結合が、ペプチドまたは疑似ペプチドによって模倣される構造モチーフを与える場合、例えば実際のHIVのV3ループと、その結合を競うような生物学的機能に有効なトポロジカルなおよび静電的な性質に加えて、ペプチドまたは疑似ペプチドは、むしろこのようなA18／Q31／H33アミノ酸残基の空間配位を有する。

また、一致検出法により検出された、共変するk項数に基づいて設計されたペプチドまたは疑似ペプチドを、抗原として使用することが可能である。すなわち、分子が模倣する生物学的機能は、動物内の免疫反応を引き起こしている。同様に、本明細書に記載の共変するk項数を含むワクチンも、本発明に含まれる。

モーガンおよびその共同研究者ら(Morgan et al. 1989. In Annual Reports in Medicinal Chemistry. Ed.:Vinick, F. J. Academic Press, San Diego, CA, pp. 243-252)は、ペプチド疑似物を「レセプターおよび酵素に相互作用して、ペプチドの適切な代用品としてはたらく構造」と定義している。この疑似ペプチドは親和性だけでなく効用および基質機能を保持しなければならない。本開示の目的のために「peptide mimetic（ペプチド疑似物）」と「peptidomimetic（疑似ペプチド）」は、上記の抜粋された定義に従って、交換可能に使用される。すなわち、疑似ペプチドは、その構造を限定されることなしに、ある特定のペプチドの機能を示す。例えば、上記に仮定されたV3ループの構造モチーフの類似物のような本発明の疑似ペプチドには、望ましい機能的特徴を提供するアミノ酸残基またはその他の化学的部分を含むこともある。

さらに本発明は、本発明の一致検出法を用いて同定された構造モチーフを有するタンパク質に相互作用するリガンド、リガンドを含む薬学的組成物、および薬学的に許容されるそれらの担体または賦形剤を提供する。リガンドには、適切な同一性を持ち、モチーフの対応する残基または一部分とその部分が相互作用するようにお互いに空間的に配位された、化学的部分が含まれる可能性もある。モチーフとの相互作用により、リガンドはモチーフを含むタンパク質のその領域の機能に支障をきたすかもしれない。

従って、本発明はヒト免疫不全ウイルス（HIV）のエンベロープタンパク質と相互作用するための薬学的組成物を提供し、これにはA/18/Q31/H33残基の空間配位を有するV3ループの構造モチーフと相互作用する官能基をもつリガンド、および薬学的に許容されるそれらの担体または賦形剤を含む。リガンドは、例えば、結合する能力があり、残基18に結合するためにリガンド内の有効な位置に存在する1番目の官能基、結合する能力があり、残基31に結合するためにリガンド内の有効な位置に存在する2番目の官能基、および、結合する能力があり、残基33に結合するためにリガンド内の有効な位置に存在する３番目の官能基等のような、モチーフと相互作用する複数の官能基を有することもある。

さらに本発明は、例えばヒト免疫不全ウイルス（HIV）のエンベロープタンパク質のようなタンパク質の構造モチーフと相互作用させるためのリガンドの設計方法を提供する。例えば、モチーフが、前述の一致検出法により同定されたA/18/Q31/H33モチーフに関係する可能性がある場合、設計方法にはHIVエンベロープタンパク質のV3ループ内のA18、Q31、およびH33の空間配位を有する鋳型を提供し、設計されるリガンドがモチーフに結合するための少なくとも一つの有効な官能基を有するような空間の制限をともなう有効なアルゴリズムを用いて、化学的リガンドをコンピューターで展開する手順が含まれる。提供された鋳型にはさらに、ト

ポロジカルなおよび／または静電的な特徴が含まれ、有効なアルゴリズムにはトポロジカルなおよび／または静電的な制限が含まれる可能性がある。同様な方法の手順には、一致検出法により同定されたモチーフを有する他のタンパク質につ

いても使用されることがある。

また本発明は、タンパク質の構造モチーフに結合するリガンドを同定する方法を提供する。構造モチーフは好ましくは一致検出法により同定される。例えば、上記のHIVエンベロープタンパク質の残基A18、Q31、およびH33を有する一致検出法によって同定されたモチーフの場合、この方法は、HIVエンベロープタンパク質のV3ループ内のA18、Q31、およびH33の空間配位を有する鋳型を提供する手順、分子の構造および定位を含むデータベースを提供する手順、およびモチーフと相互作用するように相互に配位された有効な部分を含んでいるかを決定するためにデータベース上の分子をスクリーニングする手順を含む。データベースはさらに、分子のトポロジカルなおよび／または静電的な特徴を含み、スクリーニングの手順はさらに、モチーフと相互作用するために効果的であるかどうかを決定する手順を含むことがある。例えば、データベース上に記載された分子は、それが、残基18と相互作用する第一番目の部分、残基31と相互作用する第二番目の部分、そして残基33と相互作用する第三番目の部分を有するという物理的／化学的特徴を有する可能性がある。同様の方法の手順は、所期の構造モチーフを有するその他のタンパク質に使われることもある。

本発明により提供されるリガンドが薬学的組成物に含まれる場合には、薬学的組成物はさらに、薬学的組成物に関する分野の当業者に周知の、薬学的に許容される担体も含む。本明細書で用いられる「薬学的に許容される担体」には、塩類溶液およびバッファー水溶液などの希釈剤、ならびに固相、気相、または液相のの基剤、またリポソーム等の担体(Strejan et al., 1984. J. Neuroimmunol. 7：27)、およびグリセロール、液体ポリエチレングリコール等の分散剤、その他が含まれる。薬学的組成物には、溶媒、分散媒体、コーティング、安定促進剤、抗菌および抗真菌剤（例えば、パラベン、クロロブタノール、フェノール、アスコルビン酸、チメロサール）、等張剤（例えば、塩化ナトリウム、糖、マンニトール等のポリアルコール）、ならびに周知の吸収遅延剤（例えば、モノステアリン

酸アルミニウムゼラチン）のいずれかが含まれていてもよい。

また、生物学的標的に結合するような、本発明により提供されるリガンドは、診断的な目的のために利用されることもある。本発明による診断剤には、一致検出法により同定された構造モチーフを含むタンパク質と相互作用するリガンド、およびリガンドに結合した検出可能な標識が含まれてもよい。検出可能な標識は、例えば蛍光物質または放射性物質等の、本技術分野において周知の検出可能な物質のいずれであってもよい。また、標識は、検出可能な（例えば発色する）産物を生ずる反応を触媒する酵素（例えば、ホースラディシュペルオキシダーゼまたはアルカリホスファターゼ）であってもよく、またはそのような酵素の基質であってもよい。

記述された原理の薬物発見背景への応用

数十億ドル規模の製薬産業は、高分子（「標的」）と相互作用し、標的の構造、機能または活性をある程度抑制、促進、阻害、加速、さもなければ改変する低分子（「リガンド」）の設計または発見および精製に、大いなる基礎を置いている。疾病の機構において、ある程度示唆されるのは、標的の構造、機能、または活性である。標的分子は大抵、酵素もしくはタンパク質レセプター、もしくは核酸、またはそれらの組み合わせである。可能性を持つ多数のリガンドが存在するものの、一つまたは複数の標的とともに、または対抗して働く、すなわち、疾病に対して効果のある治療用化合物として開発され市場に出されているのは、そのうちの比較的に一部でしかない。

従って、膨大な数の、可能性のある有効な化合物を考えることができ、しかも、有用でない、安全でない、有効でない、または経済的に可能でないかもしれない可能性のある化合物に基づく治療の開発に資源を使い過ぎることを回避することを可能にすることは、バイオテクノロジーおよび薬学の研究者にとって非常に関心のあることである。本明細書に記された方法は、良質で効果的な化合物を発見する方法および公共もしくは民間の分子コレクションまたはコンピューターのデータベース画像上の化合物のうち、見込のある化合物を、見込のないまたは見込の少ない化合物と区別する方法を促進および加速するために用いることができる。これらの方法は、本明細書の様々な方法において標的の構造を理解し推論す

ることを手助けし、またその幾何学的な、トポロジカルな、静電的な、またはその他の特徴をもつために、標的と効果的に相互作用する候補となるリガンドを検出することにより、効果的に利用されて価値を提供することができる。

本明細書により記載された原理の、分子およびそれらの特徴のデータベースへの応用

コンピューターデータベース上の多くの分子構造（主要メモリー内、磁気ディスク、テープ、またはその他の電子的もしくは光学的メディアにより貯蔵された）を表記する方法の一つに、「スクリーン」によるものがある。当業者らはあるスクリーンや属性が、例えば、硫酸基のような、ある特定の下部構造パターンの存在または不在をあらわす、二進の(binary)属性としてスクリーンを認識する。もし化合物の一集合が、スクリーンで表記される場合、「C」で表記されるある特定の化合物は、1および0の連続により表すことができ、1はCが含むあらかじめ決められたの下部構造パターンを表し、0はCが含まないあらかじめ決められた下部構造パターンを表す。

この方式はまた、本明細書の別の箇所で説明するように、属性による、核酸またはタンパク質に第一次構造の表記にも用いることができる。第一次構造は、「配列」、すなわち、DNAまたはRNA中の塩基またはヌクレオチドの配列として、またタンパク質中のアミノ酸残基とも呼ばれるアミノ酸配列として知られている。例えば、天然に存在する20の標準アミノ酸の一つひとつに対応するアルファベットの文字である記号を用いた記号の配列として、タンパク質の配列を表記することは容易である。また、この表記を20の二進の属性の集合よる配列において、個々の残基または位置を表記して、もしそのような表記が望ましい場合に、この表記を変換することも容易である。これらの属性は、上記のスクリーンのようにはたらく。例えば、タンパク質Pの最初のアミノ酸が、Aで表されるアラニンであれば、「位置1のアミノ酸はアラニンか」という問いを表す属性において、「1」の値により表記することができ、また、「位置1のアミノ酸はシステインか」、「位置1のアミノ酸はフェニルアラニンか」等を表す属性において、「0」の値により表記することができる。図15はアミノ酸および残基の位置を表す。

また、属性の用語を用いて、化合物のその他の側面または性質を表すことも容

易でわかりやすい。例えば、ある化合物Cが、特定の標的Tに対して活性があることが知られていれば、この場合「Tに対して活性か」という問いに対応する属性は、化合物Cに対応する対象について、1の値をとる。また、その他の例としては、製薬会社は、一連の「測定」や生物学的または化学的活性の検定を行って、多くの化合物を管理している。測定は、例えば、標的に対する有効性、血液脳関門を通過する能力、または有毒性などに関する側面を検査することもある。測定結果は、当業者に周知の前処理の方法を通して、離散値、および二進の属性の用語で表すことも可能である。特定の化合物のその他の特徴には、文献引用（すなわち、その化合物が記述、設計、発見または分析されている論文や研究の参考文献）、または化合物の所有権ならびに特許の状態が含まれ得る。

スクリーンおよびその他の属性の用語で、低分子の治療用化合物のみでなく、DNA、RNA、ペプチド、タンパク質、炭水化物および脂質等の、より高分子の、可能性のある治療用分子を表記することができる。標的分子もまた、この方法で表記できる。必要なものとしては、研究者または利用者により重要とみなされた下部構造パターンまたはその他の特徴の、あらかじめ決められた（更新、変更、縮小、または拡張されている可能性があるが）一覧表のみである。標的の構造について、下部構造パターンや、一次直線構造（配列）、遺伝的連鎖情報、疾患経路における他のタンパク質との相互作用、文献引用などの表記が望まれるかもしれない。特定の分子は時に、データベース上で、複数の対象として、つまりその分子がとり得る様々な高次構造を表す異なった対象として記録されていることもある。

化合物のデータベース表記における、スクリーンおよびその他の属性の使用は、本発明の作業を述べる上で使用されたNxMデータ行列の用語で表記することも可能であることは明らかである。NxMデータ行列は、以下の表1に示されている。

表1の行は、分子、化合物、分子構造、または配列の集合に対応し、一方列は、分子の下部構造パターン、測定結果、またはその他の側面を含む特徴に対応する。表中のセル[i, j]の数値は、もし、分子iが特徴jを持っていれば1、そうでなければ0となる。

	特徴１	特徴２	...	特徴N
分子１	1	0	...	1
分子２	0	1	...	0
...	...	...	...	...
分子M	0	0	...	1

表1

本明細書に記載の方法を分子データベースの分析に応用するための手順には、以下のものが含まれる。

1. 所期の1次元、2次元および／または3次元の分子構造の離散属性表記を支持する分子データベースを得る（または、分子データベースを得て、このような表記を作成するための標準方法を使用する）。所期の分子の配列およびその他の情報を属性表記に変換するための標準方法を使用する。

2. データベース上の個々の化合物が、具体化されたデータ行列中のM個の対象（行）の一つまたはそれ以上に対応し、個々のスクリーン表示された下部構造パターンがデータ行列の属性（列）に対応するように、このデータベースのすべてまたは一部を本発明の態様に提供する。活性、測定結果、使用された化合物に対する既知の標的、化合物の生産または貯蔵のよりどころまたは方法、化合物の所有権または特許の状態を示すなどの付加的な属性および下部構造パターン属性が共に、データ行列上のN個の属性（列）を含む。

3. 上記の基本的方法またはデータ行列上での本明細書に記載のその他の態様の一つを使用する。

4. 発見された相関のあるk項数の属性を以下のものに指向する。

グラフィカルビューアー、または、

●規則に基づくシステムのための規則作成プロセッサー、または

●利用者、研究者、もしくは管理者のためのレポート、またはレポート作成システム、または、

●データベース上に表記された化合物、配列、または構造の、ある種のさらなる

分析を行う別のコンピュータプログラム、または、

●データベース上のいくつかの変換もしくは最適化を実行するの別のコンピュータプログラム、または、

●薬物のスクリーニング実験または治療用化合物の設計、精製、もしくは生産における人間および／またはロボットを指導する別のコンピュータプログラム。

この薬物発見の応用において、本発明の結果は多くの可能な方法で利用することができる。

先ず、スクリーンに基づく分子の表記を、設定または最適化に使用することができる。例えば、相互に相関関係がなく、おおよその可能性が等しいスクリーン（属性）の集合を、良好なスクリーンに基づく表記に使用すべきことは本技術分野において公知である。本発明の方法は、上記の様に使用された場合、相関のあるスクリーンの集合を産出することがある。改変されたスクリーンの集合を、相関がなく、可能性の等しいという理想により近づけるため、スクリーンが表記する特徴を付加、除去または組み合わせるために、この情報を使用することができる。

この方法により製造される情報の、その他の便利で価値のある側面には以下のものが含まれる。

例えば、研究者が、標的の構造、標的構造の活性部位、または生物学的システムにおける数種のタンパク質のどれが標的であるかさえも知らない場合でさえ、in vivoやin vitro実験ではたらく良質の「最も重要な化合物」を製薬会社が保持することは、珍しいことではない。本明細書記載の方法が下部構造パターンと測定結果との間の相関関係を検出するために利用される場合、この情報は研究者が構造に望ましい活性を組み合わせることができるので、標的構造を示唆したり、より効果的で重要な化合物を設計したりすることの助けとなる。

また、別の例としては、本明細書で後述するように、DNA、RNANまたはタンパク質の配列の整列した集合に対応する薬物発見データベースのその箇所で、相関するアミノ酸残基をみつけることである。この場合、相関するk項数（位置）の残基は、進化上保存された構造的および機能的関係に対応することもある。従っ

て、本明細書に記載の原理はこのようにして、レセプターおよび酵素のような薬学的標的を含む、重要な生物学的高分子の構造および機能を予測または解明するこ

とにも役立てることができる。

さらに、別の例として、ひとつの標的分子であるT1の構造的または機能的側面、疾患経路またはその他の側面と、別の標的分子T2との相関を検出し、またはT1を目的とした可能性のある治療用化合物の集合の構造的、機能的、またはその他の側面と、T2を目的とした可能性のある治療用化合物の集合の構造的、機能的もしくはその他の側面との間の相関を検出することが可能である。これらの場合には、この相関関係の情報は、薬物の設計者がT1に対して有効な知識、化合物、および技術をT2に対する成果に応用することができるので、有効である。

また、本明細書記載の原理の、薬物発見および医科学へのかなり異なった応用として、上記のデータ行列変換の検討があげられる。対象（行）としての化合物および属性（列）としての化合物の特徴の代わりに、化合物が列に、そして化合物の特徴が行にそれぞれ対応する場合に可能なことを考察する。下記の表２を見ると、この方法に本発明を使用すると、特徴の空間上に相関したk項数の化合物を産出する。k項数を産出することで、数種類の貴重な情報を具体化することができる。例えば、もし行における特徴のほとんどが、下部構造パターン（スクリーン）を表している時、産出されたk項数は、化合物のクラスターに対応する。このような化合物のデータベースのクラスター化は、生物学的／化学的測定（in vivoまたはin vivo）およびコンピューター測定の両方を用いた、高処理量スクリーニング（ハイ・スループット・スクリニーング、HTS）に非常に役立つ。HTSでは、最初に個々のクラスターの一つまたは少数の群を測定し、「ヒット」が起った時のみ（すなわち、生物学的または化学的活性測定において化合物が「検出」に「合格」すること）、対応するクラスターの他の種類を分析することが、測定する際において有効で経済的である。

特徴の空間に化合物をクラスター化するために、先に示した分子データベースの「移項(transpose)」における方法の利用が、表2に示されている。分子、化合

物、分子構造または配列の集合に対応しているのは、ここでは列であり、一方、行は、分子の下部構造パターン、測定結果、またはその他の側面を含む特徴に対応している。M'個の行とN'個の列があり、ここでは、上記の元来のMおよびNに対して、おそらくM'=Nであり、N'=Mである。表のセル[j, i]中の数値は、もし、分子iが特徴jを有するなら1であり、有さない場合は0となる。

	分子１	分子２	...	分子N'
特徴１	1	0	...	1
特徴２	0	1	...	0
...	...	...	...	...
特徴M'	0	0	...	1

表２

遺伝子ネットワークを発見および分析するための本明細書記載の原理の応用

大規模なゲノム地図作成や配列決定の作業に応用される高等分子生物学的な計算技術により、完全なゲノムの配列、完全な遺伝子の発現パターン、およびこの情報を貯蔵し操作する能力に接近することができ始めている。このような情報は、新規の疾患標的および有効な治療用化合物の発見を加速させるために利用することができる。特定の物理的形質の「青写真」を形成する遺伝子および生物内のシステムは、複雑な方法で、ともに作用することが知られている。遺伝子は相互調整しながら相互作用し、それ自身および他の遺伝子の活性化ならびに発現を促進、抑制、さもなくば変調させる。

分子生物学は従来、単離した個々の遺伝子の研究に集中してきた。しかしながら、神経発達または腫瘍形成などの複雑な生物学的現象を理解するためには、例えば、一過性パターンや解剖学的パターンを計算に入れながら、同時に、数十や数百の発現パターンを研究する必要がある。このような分析は、本明細書記載の原理により提供されるような、新規の計算および統計能力を必要とする。

多くの変更が可能であり、当業者により計画されるが、遺伝子ネットワークの解析における本明細書記載の方法を利用するための基本方式には、以下の手順が

含まれる。

手順1関心の遺伝子を選択する。

手順2特定の時間での遺伝子の状態を表記するための、生物学的パラメータを選択する。生物学的パラメータには、遺伝子の発現、関連するmRNAまたはタンパク質産物の濃度、生物学的に重要なリン酸化またはその他の翻訳後修飾等のタンパク質の特定の状態、あるタンパク質の位置、または共同因子の存在または不在が含まれる。例えば、ポリメラーゼ・チェーン反応（PCR）技術を用いて増幅し、続いて周知の方法を用いて個々の遺伝子のmRNAの濃度を測定し、その後個々の遺伝子の最大の発現レベルによって分割してこれらを標準化し、これらの継続的に変異するレベルを本明細書を通して記載のデータ行列形式で表記できるz離散レベルの集合へ量子化することができる。また、遺伝子の活性および活性間の指標として、タンパク質産物の濃度レベルを用いることも可能である。時間測定された観察中のタンパク質濃度の変化は、主に三つの方法により支配されている：ある遺伝子のタンパク質合成の、その他の遺伝子のタンパク質産物による直接的調節（特殊な場合の自己調節も含まれる）；細胞核間の分子輸送；およびタンパク質濃度の減衰である。

手順3分析中の遺伝システム上の遺伝子の生物学的パラメータを経時標本抽出する方式を選択する。個々の適切な時間に、選択した遺伝子の選択した生物学的パラメータを測定するための当技術分野において周知の方法を使用する。

手順4選択した生物学的パラメータの用語により、選択した遺伝子を表記し、データ行列上の属性として、生物学的パラメータの測定値を表記する。データ行列上の行には、過時抽出標本（生物学的パラメータの測定の例）を表記する。すなわち、データ行列およびi番目の行のj番目の列のセルに対しては、i番目の経時抽出標本での、j番目の生物学的パラメータに対する（これは、j番目の遺伝子に対応する場合もしない場合もあり、それぞれの遺伝子について一つまたは複数のパラメータが測定されているかどうかによる）測定された量または特徴を入力する。記録された量、レベル、または特徴は二進（例えば、遺伝子が「オン」または「オフ」）であるか、z離散量の一つであることもある。本明細書の別の箇所

に記載したように、すべての離散量の属性はその値がある対象内に存在するか不在かの二進暗号化して表記することができ、そして本発明の好ましい態様は、この種のデータに応用することができる。

手順5本明細書記載の、データ行列上に上記の基本的方法または別の態様を利用する。

上記の手順の結果、すなわち、相関するk項数の属性の集合は、相関する遺伝子

の群の集合として理解することができる。例えば、ある遺伝子が「オン」で、別の遺伝子が「オン」ということが見つかるかもしれない。また、ある遺伝子G1が「低い発現」の時、別の遺伝子G2は「オフ」であったり、G1が「中程度の発現」では、G2は「低い発現」であったり、また、G1が「高い発現」では、G2は「中程度の発現」であることが発見されるかもしれない。このような結果は、Ｇ１がＧ２の発現を促進する、すなわち「Ｇ１がＧ２をオンにする」という仮定を支持するものであるかもしれない。同様に、相関のあるk項数の遺伝子または生物学的パラメータは、ある遺伝子が別の遺伝子または、別の遺伝子集合を抑制または「消す」こと等の証拠を提供するかもしれない。このような情報は全て、例えば、相互作用する遺伝子集合の「ブールネットワーク(boolean network)」のようなモデルを構築することに役立つ。このようなモデルは、疾病を診断、予防、および治療し、効果的で経済的に価値のある治療薬を設計する際の、価値ある補助を提供するものとして、当業者に周知である。

表3の行は経時抽出標本の集合（また、時間の点や、時間の一部分として知られている）、すなわち、特定の遺伝子または遺伝子産物の活性の観測の時刻または期間に対応する。列は、特定の遺伝子または遺伝子産物に対応する。表のセル[i, j]内の数値は、もし、遺伝子iが「オン」と考えられる、すなわち例えば、時間jの間に「活性である」または「発現されている」と考えられるなら1であり、そうでなければ0となる。この表記および応用は、遺伝子の単純なオン／オフの状態が、例えば遺伝子の主要なタンパク質産物の観測量等のような、発現のzの異なる量の集合に置き換わる場合にも、容易に発展させることができる。単一

の遺伝子の状態を表記するために、複数の生物学的パラメータが用いられる状況でも、容易にこの表記と応用を発展させることができる。

	遺伝子１	遺伝子２	...	遺伝子N
時間１	1	0	...	1
時間２	0	1	...	0
...	...	...	...	...
時間M	0	0	...	1

表３

本明細書記載の方法は、（G. S. Michaels, D. B. Carr, M. Askenazi, S. Furhman, X. Wen, and R. Somogyi, Pacific Symposium on Biocomputing 3:42-53, 1988）に記載のように、ラットの脊髄の発達に関与する遺伝子に対する遺伝子発現データの集合に応用されている。データ集合はこれらの著者から入手可能であり、1998年の３月付けのものは、ワールドワイドウェブ（WWW）http://rsb.info.nih.gov/mol-physiol/PNAS/GEMtable.htmlで、入手可能である。

逆転写ポリメラーゼ・チェーン反応(RT-PCR)の手法を用いて、112の遺伝子の発現（最大の発現レベルに標準化されたmRNAレベル）を、９つの発達時刻の点（E11、E13、E15、E18、E21、P0、P7、P14、およびP90または成体、Eは胚、Pは出生後を表す）で測定した。使用された遺伝子の一覧表には、九つの主要遺伝子系統群を含む中枢神経系(Central Nervous System : CNS)の発生において重要であると考えられる遺伝子が含まれる。

上記のデータ集合は、本明細書記載の二三の手順による方法で、分析に好都合な対象および属性のデータ行列へ容易に変換された。

1. 実数の（すなわち、連続的な数値の）遺伝子発現のレベルは（C. S. Wallace and D. L. Dowe, 「Intrinsic Classification by MML-the SNOB program」, Proceedings of the Seventh Australian Joint Conference on Artificial Intelligence, pp. 37-44, 1994）に記載の通り、SNOBソフトウエアーで具体化されたように、ベイジアンクラスター化法（Bayesian clustering method）を用いて、離散量の集合へと変換された。実数を量子化または離散化するベイジアンクラ

スター化法は当業者に周知である。結果を理解しやすいように、これらの六つの離散の数値は、AからFまでのアルファベットの記号の小さな集合へ変換された。2．行列の列が112の異なった遺伝子を、また行列の行が九つの異なった発生時間の点に対応するように、データ行列が準備された。

本明細書に記載の方法は、その後、変換された遺伝子データ集合の入力において数回、それぞれの回において、パラメータr（標本抽出数）値およびT（標本抽出反復数）値の異なった組み合わせを使用して実行された。この方法は、付録のAとDに記載の態様に酷似したコンピュータープログラムを使用して、このデータ

集団に応用することができる。しかしながらこの特定の態様は、タンパク質配列分析の分野への応用に適合され、HIVタンパク質のデータにおける特定の試験に適合するように幾つかのパラメータが同定されたことを意味する。入力データにパラメータ値が適合するようプログラムを改変しなければならない。

遺伝子発現データにおけるこれらの実行は、Windows'95の運行システム下で、IMB-PC変換可能コンピューター上で行われた。各実行において、観察および分析をおこなうために、結果の表を印刷した。T=100,000かつr=5での実行の結果は、付録Eとして貼付されている。研究者は、もっとも相関の高いk項数の遺伝子の上位10位、50位、1000位（または別の順位まで）だけを印刷したいと思うかもしれない。付録Eでは、上位25位を示した。

貼付した結果の印刷物では、次に述べる形式変換が使われた：

一並びまたは、複数の並びのそれぞれのグループは、一つの相関するｋ項数の遺伝子、すなわち本明細書の別の箇所に記載のように、統計学的に独立している、個々の成分属性の低い確率を示す一つのcset（一致集合）を報告する。独立の低い確率は当業者に周知であり、本明細書で先述したように、高い相関の形態である。個々のk項数に対して、k遺伝子が示され、独立の確率に対する数値が示される。（計算値は非常に小さくゼロに近いので、小数伸長(decimal expansion)は0に省略され、この数はしばしば0と表示される）。復唱するが、低い確率値は、高い相関の度合いを意味する。個々の遺伝子に対するA...Fの記号は、量子化された発現のレベルを表記し、遺伝子の内部データ集合名、そして遺伝子に対す

るより標準的で許容された名称が続く。

産出された相関のあるk項数は、先述の科学論文中の著者により報告された結果と比較することができる。この遺伝子発現データ集合上でこれらの著者の行なった方法の中には、対相互情報量分析がある。この分析では、相互情報として知られる特定の相関測定が、112の遺伝子のそれぞれの対について測定され、高い相互情報量をもつ遺伝子のグループが互いに近くになるように、図表により示された。本明細書に記載の方法は、付録E中の結果に示されるように、高い相関をもつ遺伝子の対のみでなく、3項数、4項数などを発見することができる。付録E中の結果および先に引用した科学論文の著者らの結果を試験したところ、二つの違った方

法はお互いを確証する傾向にはあるものの、本方法は大量の属性内の相関関係の検出により効果があることが示された。例えば、本発明者らの結果のいずれかの並びの試験は、その集団の中の別の遺伝子の対もまた、他の著者らの方法で高い対情報量をもつと記されているような、相関のある遺伝子の集合を検出する。

相関のあるk項数の属性は、そのk項数からの全ての可能な対もまた、相互に相関していること、またはその逆を示唆するとは限らない。従って、対および高位のk項の相関を検出できる、本明細書に記載されたような方法は、その他の応用において、遺伝子間、またはその他の属性間の重要な高位の相関を検出することに失敗する可能性のある対の方法よりも有利である。

本明細書記載の原理の、書類検索手段に使用するためのインターネット／イントラネット書類データベース上のカテゴリーの検出のための応用

トピックまたはキーワードによる書類検索は、十分な検索手段の存在を示唆しており、実際、効果的な検索アルゴリズムの開発に多くの労力が費やされてきた。これは、しかしながら、全体的な解決策の一部分でしかなく、問題には効果的な書類のカテゴリー化の戦略が必要とされる。情報理論は、書類を整理するために利用される効果的なカテゴリーまたはトピックの集合は、相関がなくおおよそ可能性が等しくあるべきであることを示している。これらのトピックスが広い変異幅をともなう可能性とともにあらわれる時、書類の検索空間はいくつかのトピ

ックにより、過度に大きくまたは過度に小さく分割される。もし、トピック間に相関が存在するなら（すなわち、ある書類内のトピックの存在の知識が、別のトピックが書類内に見つかるより大きな可能性を示唆する場合）、トピック集合の大きさを減少させることができる（カテゴリー化集団から相関のあるトピックの幾つかを除くことによって）。「等しい可能性」問題は、本明細書記載の原理の応用により取り組むことができる。この問題は容易に統計学的技術に従うが、標準の統計学的技術は高位の確率項の検出に失敗する。「脱相関(decorrelation)」の問題は、より難解で処理しにくい。次に最適なトピック集合は、結果が利用者に返される（そして書類それ自身の構造の解釈を混乱させる）前に検索手段に強いて、必要以上に多くのこのようなトピックを試験させる。検索効果の全ての増進により、より多くの利用者がシステムを利用することが可能になるなら、この

ようなシステムの開発者には、書類の効果的なカテゴリー化の欠落の余地はない。

この方法を最適または、ほぼ最適のトピック集合の減少に応用することは、本明細書の他節で本発明の作業を説明するために本発明者らが使用したNxMのデータ行列の用語で表記することができる。応用特異的な態様では、データ行列の行はデータベース上の特定の書類に対応し、列はそれらをカテゴリー化するように意図して作成されたトピック集合に対応する（表6参照）。

表6中の行は、データベース上の書類に対応し、列は書類を分類するために使用して作成されたトピックに対応する。表のセル[i, j]中の数値は、もし書類iがトピックjを言及しているなら1であり、そうでない場合0となる。

	トピック１	トピック２	...	トピックN
書類１	1	0	...	1
書類２	0	1	...	0
...	...	...	...	...
書類M	0	0	...	1

表6

書類集合を分類するために使用する、ほぼ最適のトピック集合検索へ、本発明を応用することに関係する手順には以下が含まれる。

1. 最初のトピック集合を得る。書類検索の分野は十分に確立され、そのような集合の作出のために効果的な方法論は、当業者に周知である。

2. このトピック集合およびトピック集合がカテゴリー化する書類の集合を用いてデータベースを作成する。トピック集合があるとすれば、必要なことは個々の書類が個々のトピックを言及しているか否かを決定するために試験することである。

3. データベース上の個々の書類が、態様のデータ行列上のM個の対象（行）の一つまたは複数の対応し、個々の考えられたトピックがデータ行列の一属性（列）に対応するように、データベースの全体または一部を表記する。

4. 上記の基本方法または本明細書記載の別の態様をデータ行列上で利用する。

5. 検出された相関のあるk項数の属性を以下に指向する。

●グラフィカルビューアーまたはプリンター、または、

●規則に基づくシステムのための規則作成プロセッサー、または、

●管理者またはコンピュータデータベース検索システムの利用者に対する報告書、または報告書作成システム、または、

●例えば、相関のある変数についてさらに徹底した統計学的分析を行う等（例えば、重相関(multiple regression))の、データのいくつかの種のさらなる分析を行う別のコンピュータープログラム、または、

●データベース上で変換または最適化を実行する別のコンピュータープログラム。

トピック集合内のトピック間の統計学的に有意な相関は、トピックの効果的でない最初の選択を示すこともある。本発明の方法により検出されたk項数の相関は「高い相関にあるトピック」（「脱相関のトピック」の目標に関して）および「高い確率の同時トピック」（「ほぼ可能性の等しいトピック」の目標に関して）の両者に対応している。当業者は、ともに起ることが判明したトピックスをトピック集合から除くべきか、組み合わせるべきかを決定する指針として、この応

用における相関の結果を利用することができる。この方法での応用の結果を用いて、このような書類検索手段の管理者は、利用者の質問に応じて検索すべきカテゴリーの数を減少させることで、システムの性能を高めることができる。システム性能の向上はサービスの提供者にとって、二つの利益がある。一つは、利用者の質問に対するシステムの反応時間が減少することであり、もう一つは、提供される利用者の総数が増加することである。

本明細書記載の原理のインターネットおよびイントラネットによる検索および貯蔵への応用

インテーネットおよびイントラネットによる検索手段は、利用者の質問にとって重要なサイトまたは書類を探すために、使用者が必要とする時間の長さを調べることによって、消費者観的に順位をつけることができる。利用者が探しているものをより迅速に見つけだせることを可能にする検索手段の結果を運行する、重要なアルゴリズムの改良は、その手段の有益性を改善し、より多くの利用者がその手段を利用でき、また（インターネット検索の場合は）利用者および広告主の

両方の社会にとって、また（会社間のイントラネットの場合には）利用者および経営者との両方の社会にとって、その手段がより魅力的なものとなる。以下に、インターネットまたはイントラネットシステム上で、より迅速に利用者にとって重要な情報を得、書類の貯蔵をよりよく管理する方法を提供する本明細書記載の原理の二つの利用を述べる。以下の説明と例では、インターネット／ウェブ、従って個々のウェブページおよびウェブサイトを考える場合と、検索がそれ自体のウェブサイトよりむしろ書類に対する場合の、単一の会社または研究所の情報システム内に貯蔵されたイントラネットが考えられる場合において、述べられた原理は同等に適用する。

この説明を明瞭にするために、その検索手段に周知の、ウェブページ集合内の各ページ、またはそのような書類集合内の内部イントラネットの書類はトピックにより既に分類され、トピック集団はあらかじめ固定されているものと仮定する。目標は検索手段の通常の結果を利用者に提示することであり、利用者の要求に関連することが知られているトピックスの付加リストをもつリンクの一覧表を補

足することもある。

表7の行はウェブページの集合または、内部イントラネット書類に対応し、列はトピックスに対応する。表のセル[i, j]の値は、ウェブページまたは書類iがトピックjを言及していれば1で、していなければ0となる。

	トピック１	トピック２	...	トピックN
ページ１	1	0	...	1
ページ２	0	1	...	0
...	...	...	...	...
ページM	0	0	...	1

表7

表7は、本明細書の別の箇所で定義され説明された対象および属性を表記するためのデータ行列の形式で、本明細書記載の基本的方法または他の態様が実行されるデータベースを示している。本明細書記載の態様の性質により、表に利用されたページの数は全ウェブページの総集合である必要はないことに留意のこと。態様がこの表上で実行（または利用）される時、態様は、同一の書類内で共に頻繁

に検出されるこれらのトピックスを検出するであろう。このことは、これらのトピックスがある様式で関連し、ウェブページの集合がこれらの関連を支持すると、利用者にとって興味深いかもしれないことを示している。

長所は幾つかある。これらの態様の計算費用は、データベース上の列の数に関連して直線的に比例する。この応用において、列の数はウェブページに関連するトピックの数を表わす。この数はほぼ例外なく膨大なので、当方法のこの性質は実際に長所である。さらに、ウェブページがランダムな順番で保存されている場合、態様をウェブページの総集合の、より処理しやすい部分集合上で実行することができる。これは検索手段が存在する場合、サーバー上の使用されていない時間に、連続的または同時に、これらの関連を検出する作業を、実行可能なより小規模な作業へと分割することができる。この方法により、その実行の間のどのような点でも大きな幅（k）の新しい関連を産出することができる。多くの他の「

関連を採掘する」方法は、長い実行時間内でのより後の段階で、関連する属性のより長いk項数のみを検出する。また、このアルゴリズムにより検出された関連するトピックスの一覧表が大きくなると、これらの新しい「同時トピックス」に対するリンタを選択するページを作成し保存することができる。これは、サーバーロードを減少させる可能性がある（従って、より多くの利用者がシステムにアクセスすることが可能となる。）。また、これは検出物の統計学的な重要性に制限を与えるので、どの新しいトピック指標を保存し、必要に応じて再作成するかを選択するためにこの情報を利用することができる。

ウェブページおよび書類の貯蔵ならびに検索を管理するための本明細書記載の原理の異なる応用

インターネットおよびイントラネットの検索手段は、トピックにより、ウェブページまたは書類の空間を整理することを意図とする。一般的には、頭文字（例えばアルファベットの）順番は、全くこの空間を等しく分割する見込がない。例えば、トピック「カリフォルニア」は「ノース・ダコタ」よりも、それに関連するずっと多くのページ集合を有するであろう。トピックによる（木(tree)の低い位置では副トピックスとともに）ページの木に似た簡易な貯蔵は、「カリフォルニア」を茂った木に託することになる。この状況で便利なことは、単一のトピックスのみによるよりは、ページの検索空間を分割するためのよりよい方法であることである。先にあげた例では、カリフォルニア関連のウェブページの大きな集合を、ノース・ダコタの集団の大きさに近い、より小さな集団に分割した方がよいだろう。集合を説明する単一のトピックを、同一の空間を含む一連の関連するトピックの一覧表を置き換えることで、大きな集合を小さなものに分割したいなら、トピックによるページの整理を続けることが可能である。再度例に戻ると、もし、「カリフォルニア」が、「太陽の光」、「ワイン」および「自動車」のみと強く関連しているとすれば、我々は「カリフォルニア」の木の節を「カリフォルニアと太陽の光」、「カリフォルニアとワイン」、「カリフォルニアと自動車」、または「カリフォルニアとその他」の節の集合に置き換えるであろう。木のこの部分の高さが（この場合）一つ分低くなるので、このことはまた、このペー

ジの検索と保存を速くすることを可能にする。木の全ての節に同じ方法を繰り返して行うことは、以前よりも良好なバランスを保証するための方法を提供することとなる。新しい木のバランシング機能のこの情報で唯一漏れているのは、それら自身の関連を発見することである。ここに表記された態様を先の節で説明した同じ表に応用すると、ページ集合からこの情報を抽出できる。この方法により、どのトピックが関連しているかだけでなく、データベース上の個々の関連に対する支持のレベルの指標を得ることができる。いったん、大きすぎるトピックが同定されたら、トピックをどのように分割するかを決定するために、このトピックを含むアルゴリズムにより検出された関連の一覧表を調べることができる。

木に基づく貯蔵検索技術の利用は周知で、このようは方法には、B-ツリー(B-trees)、k-Dツリー(k-D trees)、ツリー(tries)、k-Dツリー(k-D tries)、およびグリッドファイル(gridfiles)などの変型がある。木に基づく方法の代わりに、または、それに加えて、ハッシング方式を利用できる。このような方法を全て用い、応用領域上のデータの特定分布を利用して、貯蔵（主要メモリーおよびオフラインメモリー）および実施時間の両方で有効な利益が得られる。ここに記載された態様は、上記またはその他で示されるように、データ分布のよりよい理解と探索を得るために役立てることができる。

長所には、上記の最初の方法に対してあげた全てが含まれ、さらにもう一つの

重要な長所がある。もしある質問に関連するサイトの一覧表を検出する方法をすでに使っているならば、検索の木の均衡をとらせるために必要な関連の正確な一覧表がすでに完成しつつある。

<u>売り上げ分析、ダイレクトメールおよび関連のマーケティング活動への本明細書記載の原理の応用</u>

小売店、広告／マーケティング代理店、雑誌、新聞、ラジオ、テレビ、映画、やインターネットの会社、または非営利もしくは慈善団体のマーケティング管理職員は、どのような種類の人々が購買または貢献する見込みがあるかを知る必要がある。これらや他のマーケティング状況において、過去のマーケティング・キャンペーン（その他のキャンペーンや販売促進も含まれるが、「メーリング」と

いう用語をここで使用する）から、また、重要な商品やサービスの購入からの、または慈善事業への過去の貢献から（これらの全てを「製品」と呼ぶこととする）のデータを分析することは便利で価値があることである。

マーケティングの管理職員や、セールスマン、経営幹部が例えば以下のようなことを知ることが有効である。

どの製品が一緒に購買される傾向にあるのか（同一の消費者により、おそらくは同一の売買において）。

過去のどの広告キャンペーンまたはメーリングが良い反応（製品の高い売り上げ）を引き起こし、どれが引き起こさなかったか。

どの人工統計学的要因が昨年の会社の製品の大きな総消費に相関していたか。中西部の25～40歳の女性はその会社の製品を購入しているか。

このような質問に対して、消費者や売買、人工統計学的要因、過去のマーケティング・キャンペーンおよび特定の製品の売り上げについて整理されたデータベースを分析することで、取り組むことが可能である。慈善団体では、例えば「売り上げ」や「消費者」の代わりに、「貢献」や「寄贈者」が適応されるが、基本的な考え方は同じである。主な現在の計算上の問題の一つが、大きなデータベース上の変数または属性集合の中の関連（相関）を検出することである時、これらの分析作業へ、本明細書記載の原理を効果的に適応することができる。表8は、製品の消費者の購買についてのデータベースの分析への応用を示している。表9は、

購買が記録されているだけでなく、過去のマーケティング・キャンペーンにおける情報も記録されている場合を示していること以外は、表8と同様である。たとえば、住宅の地域、年齢群、収乳群、性別、職業カテゴリーや地域社会またはレジャーに関する活動への参加などの、消費者の人工統計学的属性に対応する列を新たに挿入して、これらの方式のどちらの一方も拡大することができる。

表8の行は消費者（および／または潜在的な消費者）に対応し、列は特定の消費者に購買された（1で表される）または購買されなかった（0で表される）製品に対応している。表のセルの[i，j]の値は、もし消費者iが製品jを購入したなら

1、そうでない場合は、0である。

	製品１	製品２	...	製品N
消費者１	1	0	...	1
消費者２	0	1	...	0
...	...	...	...	...
消費者M	0	0	...	1

表8

表9の行は消費者（および／または潜在的な消費者）に対応し、列は、特定の消費者に購買されたか（1で表示される）または購買されなかった（0で表示される）メーリング（またはその他のマーケティング・キャンペーン）および製品（品物またはサービス）に対応している。表のセルの[i, j]の値は、もし消費者iが製品jを購入したなら1、そうでない場合は、0である。

	ﾒｰﾘﾝｸﾞ１	...	ﾒｰﾘﾝｸﾞn1	製品１	...	製品 n2
消費者１	1	...	0	0	...	1
消費者２	1	...	1	0	...	0
...	...	...	...	...	...	...
消費者M	0	...	1	1	...	0

表9

本明細書記載の原理を、売り上げ／マーケティングのデータベースに応用することに関係する手順には以下が含まれる。

1. 上記の様に売り上げ／マーケティングのデータベースを入手する。必要な場合

は連続量の変数を離散状の変数に変換するための周知の方法を使用する。

2. データベース上の個々の消費者が、態様のデータ行列上のM個の対象（行）の一つまたは複数に対応し、各製品またはメーリングがデータ行列上の一属性（列）に対応するように、このデータベースを、総体的にまたは部分的に表示する。メーリング属性および（もしあれば）製品属性は共に、データ行列上のN個の属性（列）を形成する。

3. 上記または本明細書記載の別の態様の一つをデータ行列に適応する。

4. 検出された相関のあるk項数の属性を以下に指向する。

グラフィカルビューアーもしくはプリンター、または

●規則に基づくシステムのための規則作成プロセッサー、または、

●マーケティング担当者、雑誌／新聞の循環を指揮する担当者、セールスマン、経営者、もしくは他のコンピューターデータベースの質問システムの利用者に対する報告書、または報告書製作のシステム、または、

●例えば、相関する変数に対して、さらに徹底した統計学的分析を行う等（例えば、重相関）のデータのさらなる分析の幾つかを行う別のコンピュータープログラム、または、

●データベース上で変換もしくは最適化を実行する別のコンピュータープログラム。

この適用の結果は、幾つかの可能な場面に役立つ。

例えば、同一の売買または同一の消費者による異なる売買のいずれかにおいて、一緒に購入される傾向にある製品の集合を有する、k項数の相関が結果に含まれる可能性がある。このような情報は、例えば、NBAバスケットボールの入場券の購入者に、NBAチームのシャツや、バスケットシューズや他の関連する商品の割引のクーポン券が与えられるなどの、「抱き合わせ販売」や共同マーケティング・キャンペーンを開発するために利用できる。バスケットボールのファンがNBAチームのシャツを着用することを好むことは、おそらく驚くべきことではないので、上記の手順により明瞭でない製品間のその他の関連を検出することができる。

また、別の例としては、特定の商品購買と相関する特定の広告キャンペーンを表すk項数の相関が結果に含まれるかもしれない。このような情報により、マーケ

ティングの管理職員が、売り上げをもっとも増加させそうな種類の新しいマーケティング・キャンペーンに、彼等の資金を集中させることを助長することができる。

消費者のデータのクラスター化における本明細書記載の原理の利用

本明細書に記載された原理のマーケティング実践へのその他の応用には、上記のデータ行列の移項(transpose)が考えられる。対象（行）としての消費者と、属性（列）としての製品や人口統計学的要素の代わりに、消費者が列に対応し、製品や人口統計学的要素が行に対応する場合に可能なものを考える（表10参照）。この方法で本明細書に記載の原理を利用すると、人口統計学的要素または購買パターンの特徴の空間に、相関するk項数の消費者または消費者の人物概評が得られる。これは、購買の習慣やライフスタイルにおおよそ類似する、消費者または消費者の人物概評のグループへ、消費者のデータをクラスター化する形態であるとみなす。このようなクラスター化は、マーケティング資金をより最適に位置付けるために、特別な「標的グループ」を設計することに役立つ。いったん、このデータの移項(transpose)が計画されると、マーケティング活動の対しての上記記載の説明に、その他の手順を全く同様に応用することができる。

消費者をクラスター化するための、先に示したマーケティング・データベースの「移項(transpose)」の方法の利用を表10に示す。消費者の集合に対応するのは、この場合は列で、行は購入された製品や人口統計学的(demographic)特徴に対応する。M'個の行およびN'個の列があり、ここではおそらく上記の元来のMおよびNに対して、M'=NでN'=Mである。表のセル[i, j]の値は、もし、消費者iが製品jを購入した、または人口統計学的特徴jを有する時は1となり、そうでない場合は0となる。

	消費者１	消費者２	...	消費者N’
製品/人口統計学１	1	0	...	1
製品/人口統計学２	0	1	...	0
...	...	...	...	...
製品/人口統計学M’	0	0	...	1

表10

医療、疫学および／または公衆衛生のデータベースの分析への本明細書記載の原理の応用

医学研究者や開業者の間では、多くのヒトの肉体的および精神的疾患や不全は

、多くの潜在的な要因因子間の複雑な相互作用により引き起こされることが知られている。このような因子には、特定の遺伝状況や、異常、生物学的病原体への暴露、食餌面、環境（空気、水、騒音、汚染）、家庭や職場での危険物への暴露、感情的ストレス、物質中毒、貧困が含まれる。ある状況の真の「原因」は、いくつかの例を説明しようとした多くの民族的および逸話的証拠はあるものの、しばしば確認が不可能なままである。健康への脅威を発見および予防する問題は、研究者や保険会社の代理店、疫学者や公衆衛生官吏らが、生存のまたは死去した、健康なまたは疾病を煩った実在の人々の大量のデータを収集し分析することにより近年、解決への手助けがなされている。データベースへのコンピューターと統計学的分析の適用において、だれもが莫大な数の変数とそれらの潜在的な相互作用の指数学的複雑さを伴う分野で戦わねばならない。この種の分析は数十、数百、または数千の変数間の相関および関連を能率的に検出する方法により大幅に改善できる。本明細書に記載の原理はこのような状況に適応できる。

医療データベースへの適応は、本明細書の他節で利用した縦Mに横Nのデータ行列の形で表記することができる。適用の特異的な態様の一つでは、データ行列の行は、衛生研究における特定の患者または被験者に対応し、列はある疾患または疾患の集合の一因であると考えられる因子に対応する。これらの因子は、復唱となるが、社会経済的要因、ライフスタイル（運動、食餌）、患者の家庭や職場環境の側面、（例えば、発ガン化学物質への暴露）、過去の治療などが含まれる（表11参照）。

表11の行は、ある研究における患者または被験者に対応し、列は潜在的な疾患因子に対応する。表のセル[i，j]の値は、もし患者iが因子jを経験または、因子jに暴露した場合は、1となり、そうでなければ0となる。

	因子１	因子２	...	因子N
患者１	1	0	...	1
患者２	0	1	...	0
...	...	...	...	...
患者M	0	0	...	1

表11

応用の特異的な態様では、暗黙の内に表記される単一の疾患のみでなく、上記のように、また表11に示された因子を伴う属性として表記される多くの異なった疾患がある可能性がある。例えば、特定の患者pが肺がんを患ってはいるが、糖尿病や心臓疾患を患っていない場合、行pは、肺がんに対応する列には1を、糖尿病および心臓疾患の対応する列には0の値を有する。

本発明を医療／疫学／ライフスタイル要因のデータベースへの適応に関する手順には以下を含まれる。

1．上記の医療／疫学／ライフスタイル／要因のデータベースを入手する。必要があれば、連続量の変数を離散状の変数に変換するために周知の方法を使用する。

2．データベース上の個々の医療／疫学／ライフスタイル要因が、態様のデータ行列上のM個の対象（行）の一つまたは複数に対応し、個々の潜在的疾病因子がデータ行列上の一属性（列）に対応するように、このデータベースを総体的または部分的に表示する。異なった疾患を表す付加の属性は疾患因子と共にデータ行列上のN属性（列）を形成する。

3．上記または本明細書記載の別の態様の一つをデータ行列に適応する。

4．検出された相関のあるk項数の属性を以下に指向する。

●グラフィカルビューアーもしくはプリンター、または

●規則に基づくシステムのための規則作成プロセッサー、または、

●医師、研究者、公共衛生官吏、支配人、もしくは他のコンピューター・データベース質問システムの利用者に対する報告書、または報告書製作のシステム、または、

●例えば、相関する変数に、さらに徹底的な統計学的分析を行う等（例えば、重

相関）のデータのさらなる分析の幾つかの種類を行う別のコンピュータープログラム、または、

●データベース上で変換または最適化を実行する別のコンピュータープログラム

この応用の結果は幾つかの可能な場面に役立てることができる。

例えば、単一または複数の疾病状況に関連した要因の集合を含むk項数が、結果に含まれる可能性がある。このようは情報はおそらく、さらなる統計学的分析を経て精製され、これらの特定疾病の理解、検定、および予防に大きな飛躍をもたらすであろう。

また別の例では、関連が以前には知られていなかったような、お互いに関連のある要因集団を含む相関のあるk項数が結果に含まれている可能性がある。特定の食餌と肥満、または特定の職業と高いアルコールの摂取量のような、関連するライフスタイル要因の発見は、公衆衛生政策と医療実践の改善にそれ自体役立つ。

このような検出された相関全てが、公的または民営の保険の提供者にとって、潜在的に大きな利益となる。なぜなら彼等は、彼等の保険統計の表や保険証券を、例えばライフスタイル、社会経済的、およびその他の要因に基づいて、健康や寿命の予測に反映させなければならないためである。

患者のデータのクラスター化への本明細書記載の原理の使用

公衆衛生、保険証券や実践への、本明細書に記載する原理のまた別の異なった応用は、上記のデータ行列の移項(transpose)を考案することにより可能である。対象（行）としての患者と、属性（列）としての潜在的疾患要因の代わりに、対象（行）としての患者が列に、属性が行に対応する時に可能なものを考えることができる（表12参照）。この方法による本発明の利用により、特徴の空間に、相関するk項数の患者または患者の人物概評が得られる。これは、患者のデータを、彼等のライフスタイルに関しておおよそ同様の、患者や患者の人物概要のクラスター化した形態であることがわかる。このようなクラスター化は、健康サービス、出張所計画、保険保護、または他の資源の最適な割り当てを可能にするために、患者または保険の申請者の特別な「危険性の低い」或いは「危険性の高い」タイプを設計することに役立つ。このデータの移項(transpose)が計画された

場合、医療およびその他のデータベースの分析への前述の応用におけるその他の手順は

、上に示した説明に全く同様に応用できる（表12参照）。

因子空間上の患者や証券の保持者をクラスター化するための、先に示した疾病要因データベースの「移項(transpose)」における原理の利用が、表12に示されている。患者、医療研究の被験者、または潜在的な保険証券の保持者の集合に対応するのはここでは列であり、一方行はライフスタイル要因や、社会経済的要因、職場の要因やその他を含む潜在的疾患要因に対応する。M'個の行と、N'個の列があり、ここでは前述の元来のMとNに対して、M'=NでN':Mである。表のセル[j, i]の値は、もし、患者iが、因子jを保持する、または因子jに暴露した時、1となり、そうでなければ0となる。

	患者１	患者２	...	患者N'
因子１	1	0	...	1
因子２	0	1	...	0
...	...	...	...	...
因子M'	0	0	...	1

表12

<u>複合システムにおける動作不良の原因の検出への本明細書記載の原理の応用</u>

コンピューターネットワークや工場の自動化などの複合の集積システムの管理者は、始めから、システムがもつ困難な診断の問題に直面してきた。システム上の一連の事象が（おそらく延長した時間中）、総体的にシステムの動作不良を引き起こす場合、動作不良の真の原因の診断は、ほぼ克服できない課題である。例えば、高いロードの状態下で、断続的に動作不良を起こすゲートウエイ・コンピューター上のネットワーク・インターフェイス・カードの診断は、ホスト・コンピューターの破壊を引き起こさないが、ネットワークの要求に応えるためにカードを使用する（代理人をもって）他のコンピューター上のエラーを引き起こす可能性がある。このような問題は、従来の診断技術を用いて探知することが極端に困難であるかもしれない。動作不良を招く総体的なシステム上の状況の、よりよ

い分析を管理者に提供する手段は、重要な問題の診断と修理を加速させるかもしれない。

本明細書に表記された原理が応用されるデータベースは、定義される必要性がある。

データベースは総体的に、時間にわたって一連の成分の状況記録として考えることができる。このデータベースの列は、本明細書を通して利用されるデータ行列の形式上で考察される場合、一連の成分を表記し、行は時間における離散点を表す。表の値は質問の時刻での個々の成分の状態（オン、オフ、待機、エラー、その他）の暗号であることが意図されている。このようは記入の手順は当業者に周知である。

表13の行は時刻、列はシステム上の個々の成分に対応する。表のセルの[i, j]の値は、時刻iでの成分jの暗号化された状態である。

	成分１	成分２	...	成分N
時間１	1	0	...	1
時間２	0	1	...	0
...	...	...	...	...
時間M	0	0	...	1

表13

本発明の方法のシステム運転データベースの分析への応用に関係する手順には以下が含まれる。

1. 前記の通り、システム成分およびそれらの状態のデータベースを作成する。システム上の成分に対する状態の集合の選択は、システムの管理者への関心の行動、および成分そのものにより行われる。

2. データ行列上の個々の列がシステム上の成分に対応し、データ行列上の個々の行が一連の時刻に対応するような、データ行列として総体的または部分的にこのデータベースを表記する。

3. 上記の基本的方法または本明細書表記の他の態様をデータ行列に応用する。

4. 検出された相関のあるk項数の属性を以下に指向する。

●グラフィカルビューアーもしくはプリンター、または

●規則に基づくシステムのための規則作成プロセッサー、または、

●システムの管理者対する報告書もしくはレポート作成システム、または、

●例えば、相関する変数に対してさらに徹底的な統計学的分析を行う等の、データのさらなる分析の幾つかを行う別のコンピュータープログラム、または、

この応用の結果は、特異的にある作動不良とともに起ることが見られるシステム上の事象を示すために利用することができる。データベースの定式化があれば、作動不良の時刻でのシステム上の成分のみに限定されず、データベースが記録を行った全ての時刻の全ての範囲へ、作動不良状態の試験を広げる事ができる。これにより、最終的に作動不良を引き起こす成分間の難解な因果関係をこの方法で照らし出すことができる。最も容易は場合には、もし、ある成分が作動不良と相関しないことが判明したなら、システム上のこれらの成分を精査から除去するためにこの結果を用いることができる。

<u>複合システムの分析への本明細書記載の原理の応用</u>

複合システムは、幾分類似した適用の大きな系列を定義する。この検討のために、複合システムは、これらのシステムが膨大な数の相互作用する個々の成分または部分を有するために、直接的な詳細的モデル化の方法がないようなシステムとして定義される。例として（しかし、これらに限定されるものでない）、経済学、個人の行動、従業員のグループの生産性、天候パターン、国家の犯罪などがあげられる。これらの個々の場合では、これらのシステムの状態を測定するために、変数や変数の集合が用いられるように（例えば、経済学の場合では、利率、株価、およびインフレ率など）、システムを正確にモデル化する周知の方法はない。これを説明するために、複合システム上の事象は、先条件(pre-condition)、作用(action)、後条件(post-condition)の形態をとる。これらの相互作用は、作用が起る前のシステムの状態、作用そのもの、作用実施後のいくつかの時刻での結果としてのシステムの状態を表す。言い換えると、以前のシステムの摂動とその結果の集合を、システムの特徴についての情報が由来するシステムの経緯として利用できる。

本明細書記載の原理を効果的に活用する、この種の複合システムのデータベースは、ある制限に直面しなければならない。あるシステムの状態を測定するために利用される変数の集合（一般的な利用、または領域の知識から導き出せる形で）がなければならないのである。これらの変数は個々のデータベース入力の前および後の状態の部分で利用される。また、システムがそれによりかき乱される可能性のあることが知られている方法を有するシステムに応用される作用の一般的な集合がなければならない。経済学の例に戻ると、作用の集合は「財政政策」の属性の下Ｃの全ての事が含まれる可能性がある。

形式上、データベースは、ゼロまたはそれ以上の先条件の変数、ゼロまたはそれ以上の作用の変数、ゼロまたはそれ以上の後条件の変数を示す属性を有していなければならない。データベースがゼロの先および後条件の変数、およびゼロの作用の変数をとる普通の場合を除くと、考慮すべき八つの場合がある。それぞれの場合、二つの重要な解釈があることに留意のこと。例えば、先条件と作用の変数はあるが、後条件の変数がない場合を考えてみる。相関関係は二つの形で発生する。データベースそれ自体がその中に後条件の変数をもたない（また、一種類の変数のみを含む全ての相関を除去するために、報告された相関関係の集合が選抜される）、またはデータベースは実際、後条件の変数を有するものの、相関の集合そのものは全く後条件の変数を含まない形である。検討のために、前者がその場合であると仮定する。いくつかの種類の変数を含まない相関の集合を残すために、より多くの種類の変数を有するデータベース上での方法の結果が常時選抜される。

もし、データベースが一種類の変数のみ（例えば一つの作用変数または一つの後条件の変数）を含む場合は、これから生じた相関は二通りのうちの一つで解釈することができる。もし、変数が先条件または後条件の変数であれば、結果は状態の原形、すなわち共に見られる傾向にある属性の値（または変数の状態と同等）の集合を示している。天候パターンの領域からの例には、雨と低気圧がある。もし、作用変数のみがデータベース上にあるとすれば、これらの間で検出される相関は、ともに行われる傾向にある決定の集合を示している。軍部の領域では、

側面演習と攻撃は同時に見られる傾向にあることが検出されるかもしれない。これらの種のデータベースは、本明細書のどこかで説明された他のデータベースと非常に似ているので（これらの場合におけるこの方法の応用であるかもしれないので）、この節はこれらを明確には示さない。

三種類の変数のうち、二種類のみを含むデータベースの場合は、総数では三つである。

先条件と作用の変数のみを含むデータベース上で見つかった相関は、領域内の状態と作用の選択の間の関係を示す。フットボールのプレイコーリング(play-calling)が一例である（これはまたどのような直接的な詳細な方法、つまりプレイコーラー(play-caller)でモデル化できない、複合システムに関連していることに留意）。ここで、相関は例えばコーチやクォーターバックのような作用を起こす本体の傾向を示している。

もし、データベースが作用と後条件の変数のみを含む場合は、検出された相関は先条件にかかわりなく、作用の集合の有効性を明瞭にする。フットボールの例に戻ると、この種の相関は当のチームがある作用を実践できるかを示す（（四回挑戦できるうちの）三回目で到達すべき長いヤーデッジは、往々にして残念な結果となり、次の四つ目のダウンで、貧困な後条件の集合につながる傾向にあるなら、チームはこの状態で無力である傾向にあることか分かるかもしれない）。別の重要な例は、薬物の相互作用である。この場合、作用は投与された薬物であり、後条件は幾人かの患者の間で報告された副作用である。

データベースが先および後条件の変数を含む場合の効用が、最初の試験で不明解であれば、これはおそらく最も右効な場合の一つだろう。ここでは意志決定者によってとられた作用に関係なく、ある領域上の状態の後に起こる傾向がある事に我々は興味をもつか、またはとることのできる作用のない（またはシステムそのものに影響する何もない）領域上に我々がいるかのどちらかである。前者の例は、フットボールでの先条件「三回目の長い」は、後条件「四番目の長い」が続く傾向があるという事実である。実際、もっとも興味深いのは後者の場合である。天候パターンの場合を考えてみよう。もし我々が後条件の「竜巻き」にしぼる

なら（すなわち、後条件で「竜巻き」の出現に関連する相関のみを相関の集合が含むように、結果の相関の集合を選択する）、これらの相関が述べることは、竜巻きが内在する前兆である。

最後の場合は最も一般的である。データベースが三つ全ての変数を含む場合である。この形態のデータベースは全ての先の種の属性の相関を含むことができることに留意。領域の例は既にあげられている（経済学、ある母集団のなかの犯罪等）。ここでは、相関は、後条件の質に基づいて作用の集団（ある先条件のいくつかの集団）を順序づけるものとして考えることができる。

最後の考察は、データベース本体が含むデータの種類である。二進値をとる属性は、本明細書全体を通じて述べられているように、この方法に容易に受け入れられる。他の種類は、離散量の範囲に限られる。この場合でない時（例えば実数または整数値の属性）、これらの値の範囲をより処理しやすい数にまで減少させるために、当の値において、変換がなされなければならない。このための好ましい方法の中には様々なクラスター化があり、当業者に周知である。

全ての場合で、この方法によって出された相関は事象に基づいた推論パッケージへの理想的な入力である。システムの状態（例えば現在の状態）がある場合、事象に基づいた推論の手段は、このシステムに応用できる作用の集合からの選択の可能な結果の分析の基礎として、本明細書記載の原理により検出された関連を利用することができる。

一般的には、本明細書にあげる原理は、意志決定者を補助するための手段として用いることができる。意志決定者は「実在」または人工である（すなわちこの方法は、興味の領域上で意志決定をすることが目的である、人工知能的手段の一部として使うことができる）。

先条件変数および作用変数を伴うデータベースへの本明細書記載の原理の応用の説明

データベースの形態に前述の制限がある場合、本明細書の別の箇所で記載された態様の応用に対する入力の必要条件が適合することは明白である。本明細書の別の箇所で引用された便利なデータ行列上で、この文脈内のM個の行は先条件お

よびとられた作用の選択された全ての集合である。もし作用を適応する本体が、かなり擬人化されるなら、これらの行はこの本体によりなされた決定と、決定がなされた時点でのシステムの状態の経緯を表す。N個の列は、システムの状況と、システムがかき乱される方法を説明する、全ての適用可能な作用の変数の集合を有している（表14参照）。

表14の行は、その状態に対応してとられた作用へと引き続く、システム状態の例または組み合わせ（システムの先条件）に対応し、列は、システムの状態とシステムに適応ができる可能な作用を説明すると考えられる変数に対応している。

表のセル[i, p]の値は、もし列pが先条件の列である時、事象における状態の変数pの測定値の暗号であり、もし列pが作用の列である時は、事象iでとられた作用の暗号である。

	先条件１	...	先条件j	作用１	...	作用k
行１	C(1,1)	...	C(1,j)	A(1,j+1)	...	A(1,j+k)
行２	C(2,1)	...	C(2,j)	A(2,j+2)	...	A(2,j+k)
...	...	...	...	...	...	...
行M	C(m,1)	...	C(m,j)	A(m,j+2)	...	A(m,j+k)

表14

本明細書の別の箇所に記載された原理をある領域に適用する前に、考えなければならない幾つかの点がある。状態の変数の集合は定義されなければならない。これは、領域そのものの当業者（例えばフットボールのコーチ、軍事アナリスト等）に任される。

先にあげられた例は、コーチによるフットボールのプレイコーリング(play-calling)と、陸軍大将（general）による軍事的決定の場合である。一般的に、本発明の好ましい実践には、作用を起こしている本体についての情報を抽出するため、この形式のデータベース上の本発明の方法が使われる。相関する状態の変数および作用はこの本体の傾向を示す。前述のように、このシステムの状態があるとすれば、本体がとりそうな決定のよりよい状況を得るために、事象に基づく推論の手段を用いて、これらはさらに分析されるかもしれない。

この種のデータベース上での本発明の使用には、税金の収集における不正行為の指標を見つけることがある。ここでは先条件を税金還付の顕著な詳細（総収入、個人や会社により報告された総税金負担、請求された税金免除など）の獲得を目的とする属性の集合とし、可能な脱税方法の集合を定義するための作用の変数を選択する。本発明により検出された相関は税金還付の種類と脱税の種類の間の関連を示す。一致検出が出された相関を統計学的に束縛するので、脱税の指標だけでなく、これらの発見の確からしさも検出できる。税金収集代理店が、彼等に送られてきた全ての税金還付を調査することができないとすると、この方法を用い、不正行為を発見する（そして政府に対してより多くの貨幣還付）結果にもっともなりようなこれらの還付の十分に選択された部分集合を見つけることができる。

紹介されるこの様な利用の最後は、保険詐欺の領域上においてであり、本明細書記載の原理の税金収集への応用に非常に類似している。先条件の変数は、詐欺の可能な指標（請求額、保険のかけられた本体に関する詳細など）であると考えられる、保険の請求における詳細の集合を獲得することを意図し、作用変数は詐欺の種類を表す。本明細書記載の原理を適応して検出された結果は、保険請求の詳細と詐欺の間の相関を示す。保険会社は、送られてきた全ての請求を調査できないので、本明細書記載の原理を応用することで、このような請求の総一覧表をより有効な調査の対象となりそうな集合へと縮小する事が可能である。

先条件および作用変数を含むデータベースの分析への本発明の方法の応用に関する手順には以下が含まれる。

1. 前記の通り、システムの状態と作用を実行している本体が行う作用のデータベースを作成する。必要であれば、連続量の属性を離散状の属性に変換するための周知の方法を使用する。

2. 個々の状態／作用の集合がデータ行列上のN個の対象（行）の一つに対応し、個々の状態の種類の側面や作用の種類が、データ行列上の一つの属性（列）に対応するように、総体的または部分的にデータベースを表示する。

3. 上記の基本的方法または本明細書表記の他の態様をデータ行列に適用する。

4. 検出された相関のあるk項数の属性を以下に指向する。

●グラフィカルビューアーもしくはプリンター、または

●意志決定者に対する報告書もしくはレポート作成システム、または、

●意志決定のための基盤として検出された相関を用いる別のコンピュータープログラム（例えば、事象に基づく推論パッケージ(a case-based reasoning package)等）、または、

●データベース上で変換もしくは最適化を実施する別のコンピュータープログラム。

本明細書記載の原理のこの応用は、作用を実施する本体の傾向に洞察を与える、相関した状態／作用の一覧表を提供し利用する。もし例えば現在の状態などの、一つのシステムの状態のみに興味があるとすれば（またはある状態の少数の側面にのみに）、その状態を伴うある側面の集合を共有しないすべての相関の結果を選択できる。得られた集合は、興味の側面に反応してとられた作用の間の相関を表す可能性がある。作用を実施する本体の方法への得られた洞察は、次の意志決定に利用することができる。

<u>先条件の変数および後条件の変数を伴うデータベースへ応用した場合の本明細書記載の原理の説明</u>

データベースの形態において、先述の制限により本明細書の別の場所で記載された態様の入力の必要条件に従うことが強制される。この文脈上でのM個の行は先条件および後条件の例または組み合わせである（ともに見るとこれらの行を状態間のシステムの推移(transition)であると考えることができる。）N個の列は推移の前と後のシステムの状態を定義する状態の変数の集合から成り立っている（表15参照）。

表15のセル[i, j]内の値は、推移(transition)の前または後のいずれかの状態変数jの測定値の暗号である。

	先条件１	...	先条件j	後条件１	...	後条件k
行１	C(1,1)	...	C(1,j)	C(1,j+1)	...	C(1,j+k)
行２	C(2,1)	...	C(2,j)	C(2,j+1)	...	C(2,j+k)
...	...	...	...	...	...	...
行M	C(m,1)	...	C(m,j)	C(m,j+1)	...	C(m,j+k)

表15

あらゆる領域上で本発明を応用する前に、幾つかの考えるべき点がある。状態の変数の集合は定義されなければならない。これは、領域そのものにおける当業者に任される。

推移の程度を定義する時間の量の選択もまた、同等に重要である。これもまた自身の経験と、抽出したい情報の種類に基づいて決定する当業者に任される。いくつかの最小の程度が、このようはデータを集める複雑さまたはこのようなデータの有効性の限界のいずれかにより、賦課されていると仮定される。この状況において、先条件と後条件の間の時間であるこの最小の程度の複数を選択できる。

少なくとも、時間のこの隔たりはシステムがこの状態を変化させるのに十分に長くなければならない。

本発明の応用の可能性のある領域には、経済および財政政策、株式市場の予測、運動選手の人材スカウトや天候パターンが含まれる。本発明の方法の詳細を適合させるために、どのようにこれらの問題を整理するかを示すために、それぞれの簡単な説明を以下に示す。

経済および財政政策の領域においては、状態が経済指標（インフレ率や利率、住宅着工やGDP等）の集合である場合の、状態の集合のデータベースを提案する。データベースの各行は、固定した時間の量により分割された二つのこのような状態（システムの先条件および後条件）を含んでいなければならない。本発明の方法で検出された相関は、経済における循環への洞察を与える。株式市場の予測では、他に対して影響をもつと考えられる株の集合（大きなものを仮定）を提案する。固定された時間の区分が推移に対して選択されたことを復唱する。このデータベースの行は、選択した期間のこれらの株の推移を示す。本発明の結果は、

この期間中、どの株の集合が相関のある様式で「動いた」かを示す。

運動選手のスカウト（例えば、若い選手のドラフト前のプロチームによるもの等）はこの様は選択の経緯の検査を含んでいるかもしれない。データ行列の各行は個々の選手に関係する。先条件の状態は、プロのレベルでの将来の活躍を示唆するものであると考えられる統計（そしてその選手について入手可能な他の全ての情報）の選択である。後条件の状態はプロのレベルでのその選手の成功を計ることを意図した変数の集合である。本発明のより検出された相関は、チームが選択を行うための将来の成功の指標の最善の集合をみつけることに役立つ。この場合、先条件および後条件は全く同一の形式ある必要はない。状態の表示が等しいことを強制するような、意図的な制限はない。

天気予測は、本発明の非常に簡易な応用である。本明細書において選択した時間の量の程度は、利用者が検出したいと願う種の情報のみに基づく。換言すると、時間量は望まれた予測の程度を決定する。もし、１日を選んだ場合、この方法によって検出された相関は１日先の天気（現在の天気を表す個々の先条件の変数に対する値の集合があるとすると）を予測することに役立つ。もし、一週間（または一ヶ月等）を選んだ場合は、これは将来にむけての予測のどれほどが拡張するかである。

一般的に、この発明の好ましい態様では、将来の状態を予測するものとして、どのように現在のシステムの状況が働くかについての情報を抽出するためにこの形態でのデータベース上で本発明の方法を用いる。確率統計学的に連結した状態とシステム間のデータ相関があれば、効果的な予測をシステムの行動について行うことができる。

先条件と作用の変数を含むデータベースへの本発明の応用に関する手順には以下のものが含まれる。

1. システムの状態が、前述のように選択された時間量の間、状態の変数によって表記されたシステム状態間の推移のデータベースを作成する。必要があれば、いずれの連続した値の状態の変数を、状態の変数に変換するための周知の方法を使用する。

2. 状態変化に対する各状態が態様のデータ行列上のM個の対象（行）の一つに対応し、各状態の変数がデータ行列の一つの属性（列）に対応するように、総体的または部分的にこのデータベースを表示する。

3. 基本方法またはここで記載された態様をデータ行列に応用する。

4. 検出された相関のk項数の属性を以下に指向する。

●グラフィカルビューアーもしくはプリンター、または

●意志決定者に対する報告もしくはレポート作成システム、または、

●決定を行うための基盤として、検出された相関を使用する別のコンピュータープログラム（例えば、事象に基づく推論パッケージ等）、または、

●データベース上の変換もしくは最適化を実施する別のコンピュータープログラム。

作用の変数および後条件の変数を伴うデータベースへの本明細書記載の原理の応用

ここではまた、データベースの形態についての先述の規制が、本明細書の別の場所で記載された態様の入力の必要条件に従うことを強制する。この文脈中のM個の行は、作用と後条件の選択されたすべての集合である。N個の列は推移の前後で

のシステムの状態を定義する状態の変数の集合で構成されている（表16参照）。

表16の行はシステムに適応された作用、それらの結果としてのシステムの状態の観察された例または仮説的な組み合わせに対応する。列はシステムに適用できる可能な作用または個々の状態の表記の変数のいずれかに対応する。列pがデータベース上の作用の一種類に対応する場合、表16のセル[i, p]の値は、とられた作用の暗号である。もし列jが、システムの状態の幾つかの側面を示すために使われた列であるなら、表16のセル[i, j]の値は、その側面の測定値の暗号である。

	作用１	...	作用ｊ	後条件１	...	後条件ｋ
行１	A(1,1)	...	A(1,j)	C(1,j+1)	...	C(1,j+k)
行２	A(2,1)	...	A(2,j)	C(2,j+1)	...	C(2,j+k)
...	...	...	...	...	...	...
行M	A(m,1)	...	A(m,j)	C(m,j+1)	...	C(m,j+k)

表16

先の例で引用したように、この種のデータベースへの本発明の方法の応用に先立って、なされなければならない決定には、時間のある点でのシステムの状態を貯蔵するために使用される状態の変数の選択、および後条件から作用を一時的に分けるために使われる時間量の選択が含まれる。これらの選択は、応用の領域の当業者に任される。選択された時間量は、ほとんどの通常の場合、作用がシステムの状態に何らかの効果を及ぼすのに十分長くなければならない。

この発明の可能な利用には、ホッケーにおける選手管理や薬物の相互作用の研究のような大きく変化する分野などが含まれる。

本明細書の目的のため、ホッケーにおける選手管理は、これらの選手の経緯の知識があるとすると、氷上での次の出番に対する選手の選択のみに関わる。この場合の作用の変数は、次の出番に選手が選ばれるかどうかを示す二進値であり、一方、後条件の変数は、ホッケーの領域内の結果の集合を示す（その出番での相対的点数、宣言されたペナルティー（罰則）、ペナルティーの長さ、放たれたショットの相対数など）。問題の定式化により、本発明により検出された発見が、選抜された選手と次の出番での結果の集合の間の相関を示すことは明らかである。前もって、敵の選手がわかっている場合には、これらの選手を作用の変数に付

け加えることができる。この場合、味方と敵チーム双方の選手と結果の集合間の相関が得られる。この知識がある場合、本発明は、コーチが有益な結果を非常に出しそうな選手を選抜することを補助するものとして役立つ。

薬物相互作用の研究は、この発明に当然適合する。ここでは作用の変数を、ある患者に薬物、または、薬物の組み合わせを投与してきたかどうかを示す二進値とする。後条件の変数は、患者により報告された副作用の一覧表を示す。本発明

により検出された結果は、患者に与えられた薬物と副作用の集合間の、統計学的に連結した相関を示す。この様に、本発明の方法は薬物使用上の配合禁忌を決定するために使用することができるが、おそらくは後続の研究が重点を置く相互作用の集合の選抜の方法として最適である。

作用と後条件の変数を含むデータベースへの本研究の応用に関する手順には以下が含まれる。

1．前記の通り、選択した時間量の間のシステムの状態と作用との間の推移のデータベースを作成し、そこではシステムの状態は状態の変数の値により表記され、作用は作用の種類の値により表記されるものとする。必要があれば、連続した値の状態の変数、および作用の種類を離散状の変数、ならびに作用の種類に変換するための周知の方法を用いる。

2．各作用集合／状態集合対が態様のデータ行列上のM個の対象（行）の一つに対応し、各状態の変数がデータ行列の一つの属性（列）に対応するように、総体的または部分的にこのデータベースを表示する。

3．基本方法または本明細書に記載された態様をデータ行列に応用する。

4．検出された相関のk項数の属性を以下のものに指向する。

●グラフィカルビューアーもしくはプリンター、または

●意志決定者に対する報告もしくはレポート作成システム、または、

●決定を行うための基盤として、検出された相関を使用する別のコンピュータープログラム(例えば、事象に基づく推論パッケージ)、または、

●データベース上の変換もしくは最適化を実施する別のコンピュータープログラム。

先条件の変数、作用の変数、および後条件の変数を伴うデータベースへの本明細書記載の原理の応用の説明

ここではまた、データベースの形態についての先述の規制が、本明細書の別の場所で記載された態様の入力の必要条件に従うことを強制する。この応用におけるのM個の行は先条件、作用と後条件の選択された全ての集合である。N個の列は推移の前後でのシステムの状態を定義する状態の変数の集合、および暗号化され

た作用の種類で構成されている（表17参照）。

表17の行は、先条件、とられた作用、および結果として得られた後条件の例、または組み合わせを示している。列は、領域上の可能な作用の種類および領域上のある状態に対する側面に対応する（先および後条件の列の両方に対して）。もし、列pがデータベース上の作用の一種類に対応する場合、表17のセル[j，p]の値は、とられた作用の暗号である。もし列pが、先条件または後条件のいずれかの側面を特定するために使われるなら、表のセル[j，i]の値は、その側面の測定値の暗号である。

	先条件１	..	先条件ｉ	作用１	..	作用ｊ	後条件１	..	後条件ｎ
行１	C(1,1)	..	C(1,i)	A(1,i+1)	..	A(1,i+j)	C(1,i+j+1)	..	C(1,i+j+n)
行２	C(2,1)	..	C(2,i)	A(2,i+1)	..	A(2,i+j)	C(2,i+j+1)	..	C(2,I+j+n)
..	..	..	..	..	..	..	..	..	..
行Ｍ	C(m,1)	..	C(m,i)	A(m,i+1)	..	A(m,i+j)	C(m,i+j+1)	..	C(m,i+j+n)

表17

前述の例で述べられたように、この種のデータベースへの本発明の方法の応用に先んじてなされなければならない決定には、ある時点でのシステムの状態を貯蔵するために利用された状態変数の選択、および後条件から作用を一時的に区別するために用いられた時間量の選択が含まれる。この場合、先条件および後条件が等しい（変数の選択に関連して）必要がないことを述べるべきである。これらの選択は、応用の領域の当業者に任される。選択された時間量は、例えば作用がシステムの状態になんらかの影響を与えるだけ十分長くなければならない。

本発明の可能な利用には、経済政策、犯罪に対する行動、および軍事戦略が含まれる。

経済の状態を定義するための変数の集合(利率、インフレ、GNP等)および統治する団体の経済政策（政府債の発行および買い戻し）の一部としてとられる作用の集合があるとすれば、その形態の経済事象、つまり現存する経済状態、実施された財政政策の手段、政策決定に続く経済状態のデータベースが作成される。本発明の方法により検出された相関は、ある経済状態下で経済政策決定の有効性の尺度を提供する。このような知識は、それが決定のある集団への経緯的な援助（

または欠落した）を示すので、経済政策を決定する際に有益であるかもしれない。

同様な調子で、犯罪対策の設置に助力するための本発明の使用は地域社会の犯罪の過去の状態、実施された政策手段、およびその結果としての地域社会内の犯罪の状態のデータベースを作成することから始まる。状態の変数には、異なる種の犯罪率（強盗や自動車盗難等）、犯罪の異なった性質（例えば、ピストルが使われたかどうか）等が含まれる。この場合の作用の変数には、様々な犯罪に対して最小に判決する指針や、「スリーストライク（three-strike）」法、死刑の適用、および教育および精神的健康のための資金集めなどが含まれる。このようなデータベース上で、本発明は現存の犯罪状況、政策決定やこれらの決定の結果に関連する相関を検出することもある。これらの相関はこれらの決定を行うことに責任を持つ人々への非常に貴重な助けになることが判明することが考えられる。

意志決定者の概念は、軍事戦略の領域においての注意深い考慮を必要とする。陸軍大将の意志決定の十分な経緯を伴うデータベースを埋めるための「実績」が十分ない場合ももっともである。このような場合、好ましい実行は意志決定者の概念を拡張して、全ての類似の意志決定者を含むことができる。一例として、戦車師団を管轄する一人の陸軍大将を考えてみる。もし、陸軍大将が最近昇進したなら、同様の義務をもつこのような全ての将軍の全経緯を考えることは賢明である。この方法の使用の程度を更に拡大するため、データベースを一人の中佐の決定よりはむしろ、歩兵中佐によりなされた決定で埋めることができる。検出された相関は、陸軍大将が決定を行った際、直面した戦場の状況の測定値があるとすると、その階級の将軍の傾向を示す可能性がある。同様に、決定の集合の結果にアクセスするので、陸軍大将らがどの戦況を稚拙に扱ったかを決める立場にいるかもしれない。このような知識は、抵抗戦略の選択に極めて重大であることが判明するかもしれない。

先条件、作用および後条件の変数を含むデータベースへの本研究を応用に関する手順には以下が含まれる。

1．前記の通り、選択した時間量を包含する状態と作用のデータベースを作成す

る。必要があれば、連続した値の状態の変数および作用の種類を、離散状の変数および作用の種類に変換するための周知の方法を用いる。

2. 各状態／作用／状態の三重がデータ行列上のM個の対象（行）の一つに対応し、各状態の変数または作用の種類がデータ行列の一属性（列）に対応するように、総体的または部分的にこのデータベースを表示する。

3. 基本方法またはここで記載された態様をデータ行列に応用する。

4. 検出された相関するk項数の属性を以下のものに指向する。

●グラフィカルビューアーもしくはプリンター、または

●意志決定者に対する報告もしくはレポート作成システム、または、

●決定を行うための基盤として、検出された相関を使用する別のコンピューターープログラム(例えば、事象に基づく推論パッケージ)、または、

●データベース上の変換もしくは最適化を実施する別のコンピュータープログラム。

この説明が好ましい態様を参照して実施され、本明細書に添付のAからEの付録に続いているページは、この説明の一部を形成する付録であり、この請求により定義されるような意図および範囲内に納まる発明の原理を実施する他の態様の作成が可能であることは、当業者に理解されるであろう。

補遺A

File colmc.pl: 1/15

```
# perl version of Evan Steeg's Coincidence Detection Algorithm,
# here applied to data which comes in rows and columns of ascii
# symbols.  Used first for tests on artificial and real (HIV)
# protein sequence data.
# march 1996

##########################################################################

$tiny_num = 0.000001;

$fact[0] = 1;
$fact[1] = 1;
$fact[2] = 2;
$fact[3] = 6;
$fact[4] = 24;
$fact[5] = 120;
$fact[6] = 720;
$fact[7] = 5040;
$fact[8] = 40320;
$fact[9] = 362880;
$fact[10] = 3628800;
$fact[11] = 39916800;

sub compare
{
   if ($a < $b)
   {
      $r = -1;
   }
   elsif ($a == $b)
   {
      $r = 0;
   }
   else
   {
      $r = 1;
   }

# print "a: $a, b: $b, r: $r\n";

   return $r;
}

sub comp_aa
{
#   my ($a1, $c1, $a2, $c2, $r);
   my ($c1, $c2);

#   $a1 = substr $a, 0, 1;
   $c1 = substr $a, 1;

#   $a2 = substr $b, 0, 1;
   $c2 = substr $b, 1;

   if ($c1 < $c2)
   {
      $r = -1;
   }
   elsif ($c1 == $c2)
   {
      $r = 0;
   }
   else
   {
      $r = 1;
```

```
    }
    return $r;
}


# calc the factorial of a number.  want (n)
# for now, it's just easier and faster to hard code them into a table
sub factorial
{
   my ($n) =  @_;

# print "n: $n\n";

   if ($n >= 0 && $n <= 11 )
   {
      return $fact[$n];
   }
   else
   {
      print "ERROR: n larger than max defined factorial requested. ($n)\n";
      exit (0);
   }
}

# calc the binomial coeff.  want r (number of iterations) and h (
# observed number of bits)
sub binomial_coeff
{
   my ($r, $h) = @_;
# print "r: $r, h: $h\n";

   $rf = &factorial($r);
   $hf = &factorial($h);
   $rhf = &factorial(($r - $h));

# print "rf: $rf, hf: $hf, rhf: $rhf\n";

   return ($rf / ($hf * $rhf));
}

# calc the chernoff.  want ($observed, $expected, $r1, $T1)
sub chernoff
{
   my ($observed, $expected, $r1, $T1) = @_;

   $diff = $observed - $expected;
   $diff_sq = $diff * $diff;
   $numerator = 2.0 * (0.0 - $diff_sq);
   $denominator = $T1  * ($r1 * $r1);
   return (exp ($numerator / $denominator));
}

# calc the ith power of a number.  NOTE: this thing can only grok
# positive integer exponents larger than 0!
sub pow
{
   my ($i, $p) = @_;

   if ($p < 0 || $p != int ($p))
   {
      print "ERROR: I can only grok positive integer exponents larger than 0!\n";
      exit (0);
   }
```

```
    $a = 1.0;
    for ($n = 0; $n < $p; $n++)
    {
       $a *= $i;
    }

# print "i: $i, p: $p, a: $a\n";
    return $a;
}


# want ($r, $h, $c_element), cset and aasites assumed as global
sub prob_coincidence
{
    my ($r, $h, $c_element) = @_;
    my @elements;

    if ($r > 0)
    {
       $joint     = 1.0;
       $joint_neg = 1.0;

       @aalist = split /\|/, $c_element;
#print "c_elelment: $c_element, aalist: @aalist\n";

       foreach $aa (@aalist)
       {
          $joint     *= $aasites{$aa};
          $joint_neg *= (1.0 - $aasites{$aa});
#print "aa: $aa, joint: $joint, joint_neg: $joint_neg\n";

       }

#       $ans = &binomial_coeff($r, $h) * &pow($joint, $h) *
#              &pow($joint_neg, ($r - $h));
       $ans = &binomial_coeff($r, $h) * ($joint ** $h) *
              ($joint_neg ** ($r - $h));
    }
    else
    {
       return (0.0);
    }
#q print "joint: $joint, joint_neg: $joint_neg, ans: $ans\n";

    return $ans;
}

sub expected_size
{
    my ($r, $c_element) = @_;

    $sum = 0.0;

    foreach $h (1..$r)
    {
       $sum += (&prob_coincidence($r, $h, $c_element) * $h);
#print "r: $r, h: $h, sum: $sum\n";
    }

    return $sum;
}


sub prob_of_correlation
{
    my ($c_element, $h_total_obs, $h_expected_total, $r, $T) = @_;
```

```
#    $h_expected_total = &expected_size($r, $c_element);
     $ch = &chernoff($h_total_obs, ($h_expected_total * $T), $r, $T);

     return $ch;
}

# randomly select a list of 'sample_size' unique sequences
# in the range from 0 to the number of rows in @family
# want sample_size, family.
sub rsample_family
{
    my $R = shift @_;
    my @family = @_;

    my (%which_rows, @sampled_family, @sampled_rows);

#    print "whichrows: ", keys %which_rows, "\n";

    # generate $R number of unique keys
    $f = scalar @family;
    while (scalar (keys %which_rows) < $R)
    {
        $n = int (rand $f);
#print "randnum: $n\n";
        $which_rows{$n} = 1;
    }
#    print "whichrows: ", keys %which_rows, "\n";

    # pick out the corresponding sequence from the 'family list'
    @sampled_rows = keys %which_rows;
    foreach $line (@sampled_rows)
    {
        push @sampled_family, $family[$line];
    }

#print "RSAMPLE\n";
#    $i = 0;
#    foreach $line (@sampled_family)
#    {
#        print $line, " : ";
#        $n = $sampled_rows[$i];
#        print $n, ":".$family[$n], "\n";
#        $i++;
#         print "$line\n";
#    }
#print "RSAMPLE END\n";
#exit(0);

    return @sampled_family;
}


# return the n'th column of an array
# want ($n, @array)
sub column
{
    my $n = shift @_;
    my @a = @_;
    my $col;

#print "COLUMN: $n\n";
#foreach (@a)
#{
#    print "$_\n";
#}
```

File coinc.pl: 5/ 15

```
    # go thru and append the n'th element of each row in @array to $col
    $col = "";
    foreach $line (@a)
    {
       $col = $col . substr $line, $n, 1;
    }

#   print length $col, ":", $col, "\n";
#print "COLUMN END\n";

   return $col;
}


# find all occurences of a character 'aa' in the n'th column of the
# array sampled_family
# want ($aa, $n, @sampled_family)
sub find_all
{
   my $aa = shift @_;
   my $n  = shift @_;
   my @sampled_family = @_;
   my ($bstring, $col);

#   print "FIND_ALL: $aa, $n\n";
#   print "01234567890123456789O\n";
#   foreach (@sampled_family)
#   {
#      print "$_\n";
#   }

#   print "JUMPING TO COL\n";
   $col = &column ($n, @sampled_family);
#   print "GOT: $col\n";

   $bstring = "";
   if ( (index $col, $aa) != -1) # make sure $aa is found in $col
   {
      for ($i=0; $i < length $col; $i++)
      {
         $c = substr $col, $i, 1;
         if ($c eq $aa)
         {
            $bstring = $bstring . "1";
         }
         else
         {
            $bstring = $bstring . "0";
         }
      }
   ;
   else
   {
      $bstring = "NOT_FOUND";
   }

#   print "$bstring\n";
#   print "FIND_ALL END\n";
#   exit(0);

   return $bstring;
}

# this subroutine isn't exactly the most optimal code, but....
sub mi
```

```
{
    my ($col1, $col2, $m) = @_;
    my ($s1, $s2, $row, $p1, $p2, $pj, $a1, $a2, $s, %sj, $contrib, $total);

    $s1 = column($col1, @family);
    $s2 = column($col2, @family);

#   print "col1: $col1, $s1\n";
#   print "col2: $col2, $s2\n";

# print "keys1: ", keys %sj, "\n";

    # calc the joint prob
    for $row (0..($m-1))
    {
       $a1 = substr $s1, $row, 1;
       $a2 = substr $s2, $row, 1;

       $s = $a1 . $a2;

       if (exists $sj{$s})
       {
          $sj{$s}++;
       }
       else
       {
          $sj{$s} = 1;
       }
#      print "a1: $a1, a2: $a2, s: $s\n";
    }
#      print "keys2: ", keys %sj, "\n";


    foreach $s (keys %sj)
    {
       $sj{$s} = $sj{$s} / $m;

       if ($sj{$s} < $tiny_num)
       {
          $sj{$s} = $tiny_num;
       }
#      print "$s: $sj{$s}\n";
    }

    $total = 0;
    foreach $s (keys %sj)
    {
       $a1 = substr $s, 0, 1;
       $a2 = substr $s, 1, 1;

       # find partial probs
       $a1 = $a1 . $col1;
       $a2 = $a2 . $col2;

       $pj = $sj{$s};
       $p1 =   $sites{$a1};
       $p2 = $sites{$a2};


        if ($p1 < $tiny_num)
        {
           $p1 = $tiny_num;
        }
        if ($p2 < $tiny_num)
        {
           $p2 = $tiny_num;
```

File coinc.pl: 7/ 15

```
        }
        if ($pj < $tiny_num)
        {
           $pj = $tiny_num;
        }

        $contrib = ($pj * log ($pj / ($p1 * $p2)));
        $total += $contrib;
#  print "a1: $a1, a2: $a2, s: $s, pj: $pj, p1: $p1, p2: $p2, contrib: $contrib, total: $tota

    }

    return $total;
}


sub incidence_vec
{
    my ($col, $key) = @_;

    my ($vec);

    $vec = "";
    if ( (index $col, $key) != -1)
    {
       for $i (0..((length $col) - 1))
       {
          $c = substr $col, $i, 1;
          if ($c eq $key)
          {
             $vec = $vec . "1";
          }
          else
          {
             $vec = $vec . "0";
          }
       }
    }
    else
    {
       $vec = "NOT_FOUND";
    }

    return $vec;
}


# given two columns, go through each letter in the alphabet and
# generate the incidence vector for them.  then if the results are
# non-zero, send them to mi2_real for the real computations
sub mi2
{
    my ($coll, $col2, $m) = @_;
    my ($s1, $s2, $key1, $key2, $total, $sum);

    $s1 = column($col1, @family);
    $s2 = column($col2, @family);

    $sum = 0.0;
    foreach $key1 (keys %alphabet)
    {
       $vec1 = incidence_vec ($s1, $key1);

       if ($vec1 ne "NOT_FOUND")
       {
          foreach $key2 (keys %alphabet)
```

File coinc.pl: 8/ 15

```
        {
           $vec2 = incidence_vec($s2, $key2);

           if ($vec2 ne "NOT_FOUND")
           {
              $total = mi2_real($vec1, $vec2, $m);

#          print "s1  : $s1\n";
#          print "vec1: $vec1\n";
#          print "s2  : $s2\n";
#          print "vec2: $vec2 \n";

#             if ($total > 1.0)
              {
                 printf "mi2, cols: %d, %d | key1: $key1 | key2: $key2 | total: %.9f\n", [$c
                 $sum += $total;
              }
           }
        }
     }
   }
   print "total sum: $sum\n";
}


# Given two columns (the actual string of amino acid symbols),
#  produce all combinations (pairs) of attr1, attr2, where attr1 is
#  an incidence vector for a symbol occurring in col1 and
#  likewise for attr2 from col2.  Then call mi2 on the pair
#  of incidence vectors.
# Compute mutual_info(attr1,attr2) where attri are binary incidence
#  vectors for two a1@col1, a2@col2.

sub mi2_real {
 my ($attr1, $attr2, $m) = @_;
 my ($a,$a1,$a2,$s,$p0,$p1,$pj,%hash_single1, %hash_single2,
    $total,%hash_joint);

 for $row (0..($m-1))
 {
      $a1 = substr $attr1, $row, 1;
      $a2 = substr $attr2, $row, 1;

      $s = $a1 . $a2;

#print "row: $row, a1: $a1, a2: $a2, s: $s\n";

      if (exists $hash_single1{$a1})
      {
       $hash_single1{$a1}++;
      }
      else
      {
       $hash_single1{$a1} = 1;
      }

      if (exists $hash_single2{$a2})
      {
       $hash_single2{$a2}++;
      }
      else
      {
       $hash_single2{$a2} = 1;
      }
```

File colinc.pl: 9/ 15

```
        if (exists $hash_joint{$s})
        {
         $hash_joint{$s}++;
        }
        else
        {
         $hash_joint{$s} = 1;
        }
    }


    foreach $s (keys %hash_joint)
    {
       $hash_joint{$s} = $hash_joint{$s} / $m;

       if ($hash_joint{$s} < $tiny_num)
       {
          $hash_joint{$s} = $tiny_num;
       }
#print "s: $s, hj: $hash_joint{$s}\n";
    }

    foreach $a (keys %hash_single1)
    {
       $hash_single1{$a} = $hash_single1{$a} / $m;

       if ($hash_single1{$a} < $tiny_num)
       {
          $hash_single1{$a} = $tiny_num;
       }
#print "a: $a, hs1: $hash_single1{$a}\n";
    }

    foreach $a (keys %hash_single2)
    {
       $hash_single2{$a} = $hash_single2{$a} / $m;

       if ($hash_single2{$a} < $tiny_num)
       {
          $hash_single2{$a} = $tiny_num;
       }
#print "a: $a, hs2: $hash_single2{$a}\n";
    }


    foreach $s (keys %hash_joint)
    {
     $a1 = substr $s, 0, 1;
     $a2 = substr $s, 1, 1;
     $pj = $hash_joint{$s};
     $p1 = $hash_single1{$a1};
     $p2 = $hash_single2{$a2};

     if ($p1 < $tiny_num)
     {
       $p1 = $tiny_num;
     }
     if ($p2 < $tiny_num)
     {
       $p2 = $tiny_num;
     }
     if ($pj < $tiny_num)
     {
       $pj = $tiny_num;
     }
```

File coinc.pl: 10/ 15

```
        $total += ($pj * log ($pj / ($p1 * $p2)));

    }
    return $total;

}


#####################################################################

# check to make sure a file name was given
if (scalar @ARGV != 4)
{
   print "usage: $0 data_file sample_size iterations min_freq\n";
   exit;
}

$filename    = $ARGV[0];
$sample_size = $ARGV[1];
$iterations  = $ARGV[2];
$min_freq    = $ARGV[3];

# read contents of file into array family
open (DATAFILE, $filename);
@family = <DATAFILE>;

chop @family;

# remove nial's +, -, and | delimiters
@@family = grep (!/\+/, @family); # get rid of lines beginning with '+'
#foreach (@family)                # remove all '|'s
#{
#   tr/\|//d;
#}

##@family = grep (/^\w/, @family);

#foreach (@family)
#{
#   print "$_\n";
#}

#while (length $family[(scalar @family) -1] < 1)
#{
#   print "Empty line: ", scalar @family, " deleted.\n";
#   pop @family;
#}

#$i = 0;
#foreach (@family)
#{
#   print "$i: $_\n";
#   $i++;
#}


#####################################################################
# NOW for the real stuff!

print "Sample_size: $sample_size\n";
print "Iterations : $iterations\n";
print "Min_freq   : $min_freq\n";

# construct aasite list
```

File coinc.pl: 11/15

```
$n = length $family[0];
$m = scalar @family;
foreach $row (@family)
{
   for $j (0..($n - 1))
   {
      $c = substr $row, $j, 1;
if (length $c != 1)
{
   print "BUG!!! $row, $j\n";
   exit;
}

      #print "$line:$j:$c\n";
      $i = $j; # + 1;
      $s = $c . $i;            # create aasite name

# print "c: $c, j: $j, i: $i, s: $s\n";

      if (exists $aasites{$s})
      {
         $aasites{$s}++;
      }
      else
      {
         $aasites{$s} = 1;
      }
   }
}

# figure out the alphabet
#@a = keys %aasites;
#print @a, "\n";
#foreach (@a)
#{
#      print "$_:$aasites{$_}\n";
#}

foreach $entry (keys %aasites)
{
   $c = substr $entry, 0, 1; # want the first character in each entry
#   print $c, "\n";
   $alphabet{$c} = 1;
}

print keys %alphabet, "\n";

# calc marginal probabilities for each column of aasites
foreach $key (keys %aasites)
{
   $p = $aasites{$key} / $n;
   $aasites{$key} = $p;
#   print "$key : $p\n";
}


for $col1 (0..($n-2))
{
   for $col2 (($col1 + 1)..($n-1))
   {
      $mi = &mi($col1, $col2, $m);
      print "columns: ", ($col1 + 1), " ", ($col2 + 1), " mi = $mi\n";

      $mi2 = mi2($col1, $col2, $m); # might as well do mi2 while we're here
   }
}
```

File coinc.pl: 12/15

```
#exit;

### MAIN LOOP

# seed the random number generator
#$seed = 111;
#srand ($seed);  # remove '($seed) to get seed from the system clock
srand();

# print "START MAIN LOOP\n";

for ($iter=0; $iter < $iterations; $iter++)
{
   my %BINS;

#   print "\nITERATION: $iter\n";
   print STDERR "ITERATION: $iter\n";

#   print "JUMP TO rsample_family\n";
   @sampled_family = &rsample_family ($sample_size, @family);

#   print "sample size: $sample_size\n";
#   print "     01234567890123456789O\n";
#   $i = 0;
#   foreach (@sampled_family)
#   {
#      print "$i : $_\n";
#       $i++;
#   }
#   print "rsample printed\n";


   foreach $aasite (keys %aasites)
   {
      $aa      = substr $aasite, 0, 1;
      $col_num = substr $aasite, 1;

#     print "aa: $aa, colnum: $col_num\n";

      $occurence_string = &find_all ($aa, $col_num, @sampled_family);
#      print $occurence_string, "\n";
      if ($occurence_string ne 'NOT_FOUND')
      {
#         print "FOUND occ_str: $occurence_string\n";
         if (exists $BINS{$occurence_string})
         {
            $BINS{$occurence_string} = $BINS{$occurence_string} .
                                       $aasite . "|";
         }
         else
         {
            $BINS{$occurence_string} = $aasite . "|";
         }
      }
   }

#   foreach (keys %BINS)
#   {
#      print "$_:$BINS{$_}\n";
#   }

   # sort the collision list associated with each BIN and throw away
   # entries with just one 'collision'
   foreach $bin (keys %BINS)
   {
```

File coinc.pl: 13/15

```
    my @aalist;

    $s = $BINS{$bin};
#    print $s. "\n";
    @aalist = split /\|/, $s;
#    $i = 0;
#    foreach (@aalist)
#    {
#       print"$i:$_\n";
#       $i++;
#    }

    if ( (scalar @aalist) > 1) # throw away single 'collisions'
    {
       # then sort the others
#       $sorted_aalist = join "|", sort comp_aa @aalist;
       $sorted_aalist = join "|", sort @aalist;
#       print "sorted aalist: $sorted_aalist\n";


       $BINS{$bin} = $sorted_aalist;
    }
    else
    {
#       print "chucked\n";
       delete $BINS{$bin};
    }

  }

#  print "SORTED BINS\n";
#  $z = 0;
#  foreach (keys %BINS)
#  {
#     print "$z:$_:$BINS{$_}\n";
#     $z++;
#  }

  # now we update the cset table
  foreach $bin (keys %BINS)
  {
     $count = 0;
     # sum up bin hits; sample_size should equal length of bins
     for ($i=0; $i < $sample_size; $i++)
     {
        $c = substr $bin, $i, 1;
        if ($c eq "1")
        {
           $count++;
        }
     }

     $key = $BINS{$bin};
#     print "cset key: $key\n";
     if (exists $cset{$key})
     {
        $cset{$key} += $count;
     }
     else
     {
        $cset{$key} = $count;
     }
  }

#  print "CSET\n";
#  $z = 0;
```

File coinc.pl: 14/15

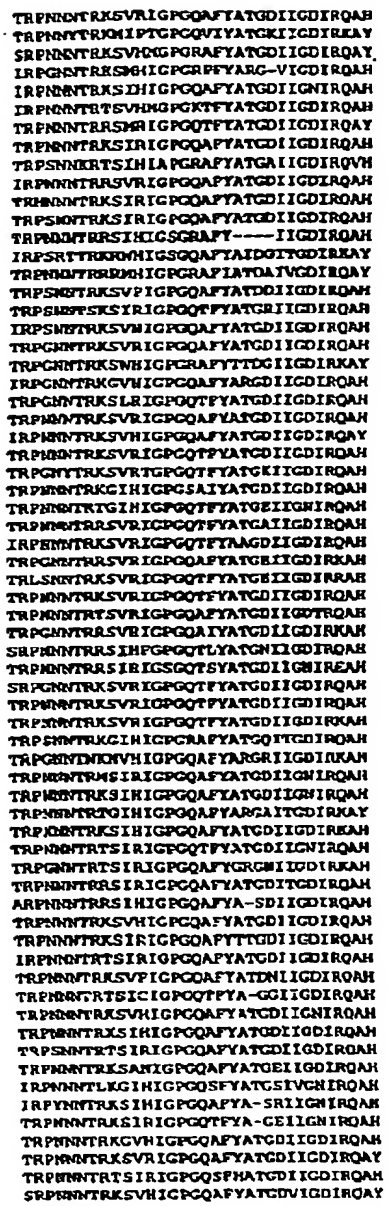
```
#    foreach (keys %cset)
#    {
#       print "$z:$_:$cset{$_}\n";
#       $z++;
#    }

#    print "$iter, BINS : ", scalar keys %BINS, "   ";
#    print "CSETS: ", scalar keys %cset, "\n";
     print STDERR "BINS : ", scalar keys %BINS, "\n";
     print STDERR "CSETS: ", scalar keys %cset, "\n";

}

print "CSETS: ", scalar keys %cset, "\n";

print "\n\nGathering stats.\n";

foreach $entry (keys %cset)
{
   $h_total_obs = $cset{$entry};
   $h_expected_total = &expected_size($sample_size, $entry);
   $correlation = &prob_of_correlation($entry, $h_total_obs,
                                       $h_expected_total,
                                       $sample_size,
                                       $iterations);
   if ($correlation < 0.000000001)
   {
      $correlation = 0.0;
   }

   if ($h_total_obs >= $min_freq)
   {
      # this is a weelly ugly hack to prevent hash key collisions
      $h = $h_total_obs;
      while (exists $output{$h})
      {
         $h = $h . "*";
      }
#      print "\nEntry      : $entry\n";
#      print "Obsrv hits: $h_total_obs\n";
#      printf "Expct hits: %.9f\n", $h_expected_total * $iterations;
#      printf "Prob corrl: %.9f\n", $correlation;
      $output{$h}[0] = $entry;
      $output{$h}[1] = $h_total_obs;
      $output{$h}[2] = $h_expected_total * $iterations;
      $output{$h}[3] = $correlation;
   }
}


@hits = keys %output;
@hits = sort compare @hits;
##@hits = sort @hits;

#foreach (@probs)
#{
#   print "$_\n";
#}

print "SORTED\n";
foreach $hit (@hits)
{
   my (@aalist);

#   $i = index $hit, "*";
#   if ($i != -1)
```

File coinc.pl: 15/15

```
#   {
#      $h = substr $hit, 0, (index $hit, ':::');
#   }
#

   $s = $output{$hit}[0];
   @aalist = split /\|/, $s;
   foreach (@aalist)
   {
     $aa = substr $_, 0, 1;
     $col_num = substr $_, 1;
     $_ = $aa . ($col_num + 1);
   }

   $s = join '|', sort comp_aa @aalist;


#  print "\nEntry      : ", $output{$hit}[0], "\n";

   $observed = $output{$hit}[1];
   $expected = $output{$hit}[2];
   $prob     = $output{$hit}[3];

   if ($expected < $observed && $prob < 0.5)
   {
      print "\nEntry      : ", $s, "\n";
      print "Obsrv hits: ", $output{$hit}[1], "\n";
      printf "Expct hits: %.9f\n", $output{$hit}[2];
      printf "Prob corr1: %.9f\n", $output{$hit}[3];
   }
}
```

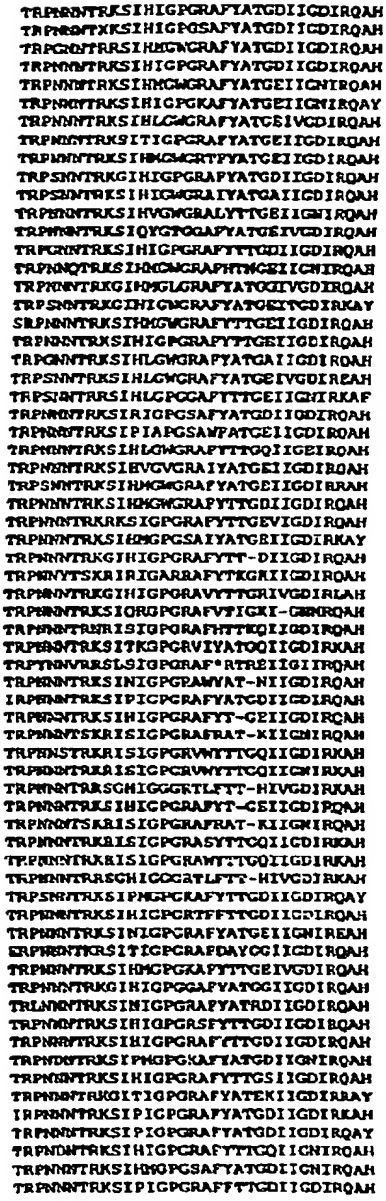
補遺B

HIV Input: 1/10

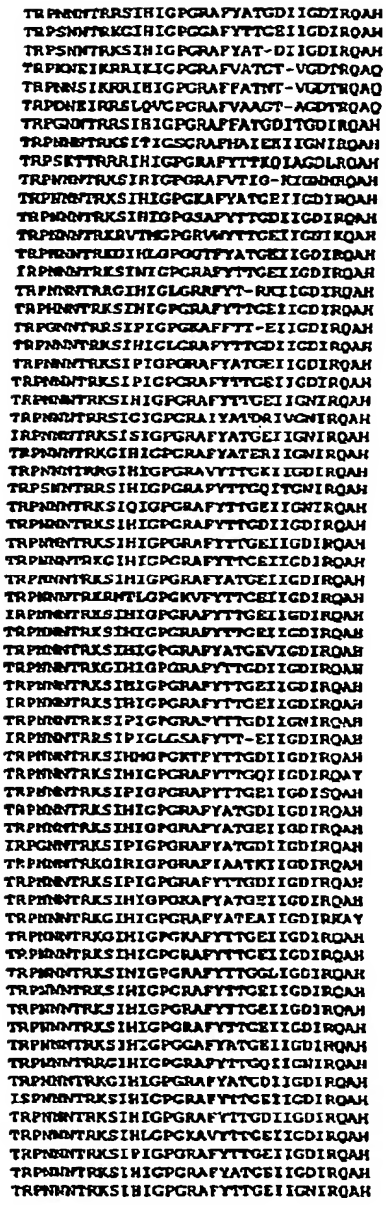


HIV input: 2/ 10

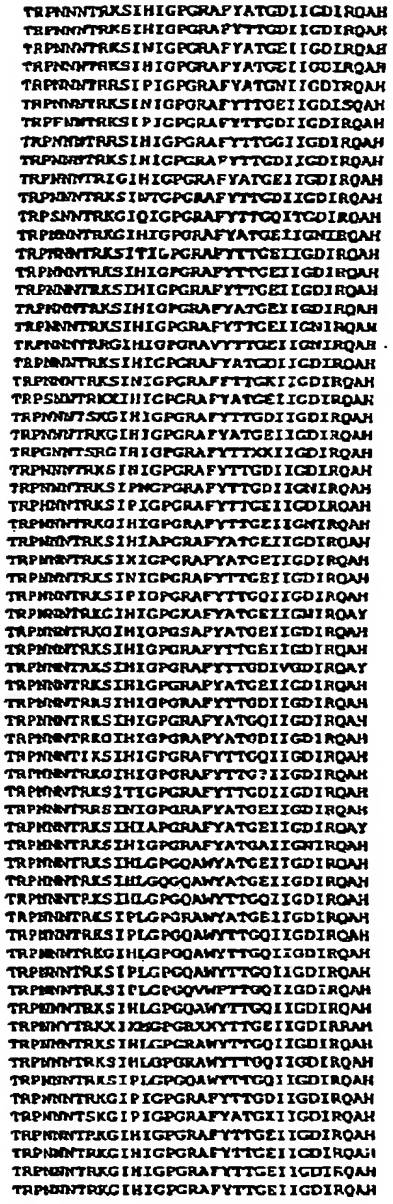
HIV input: 3/ 10



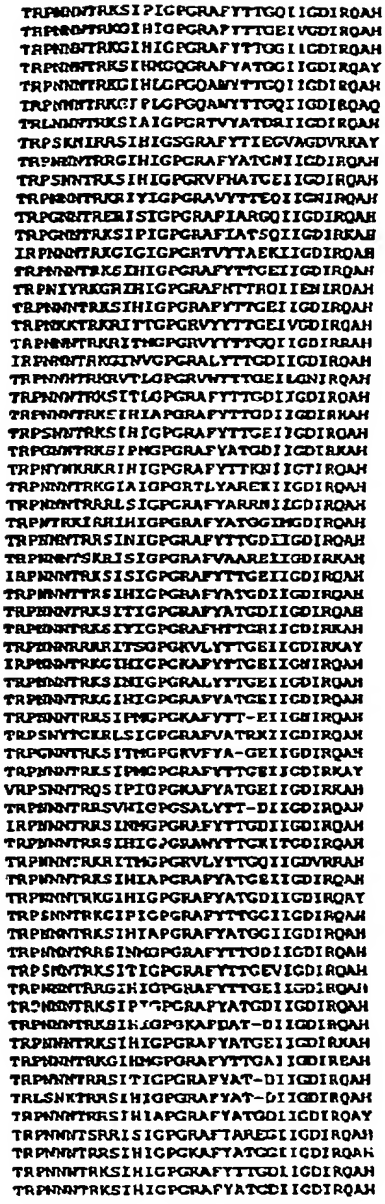
HIV input: 4/10



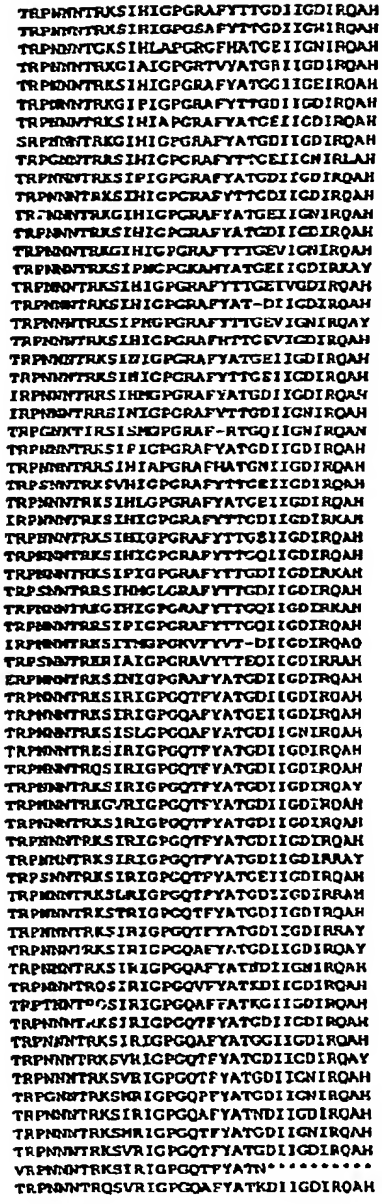
HIV input: 5/10



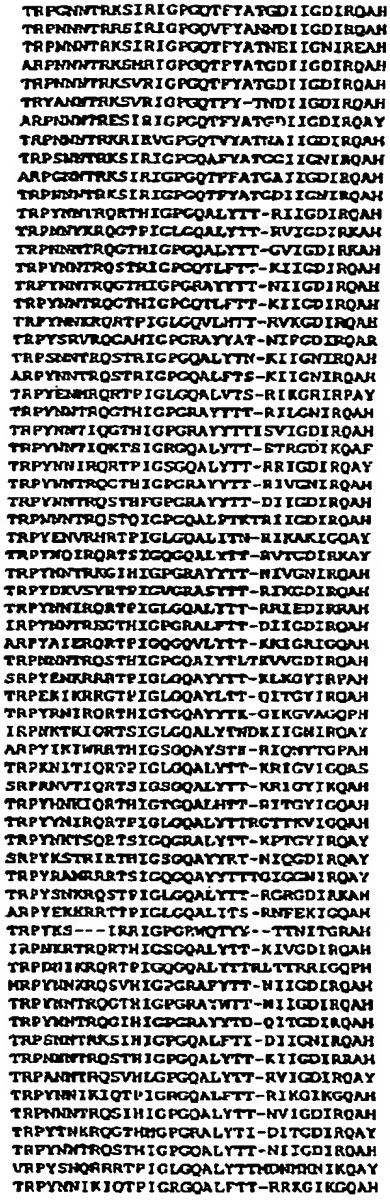
HIV input: 6/10



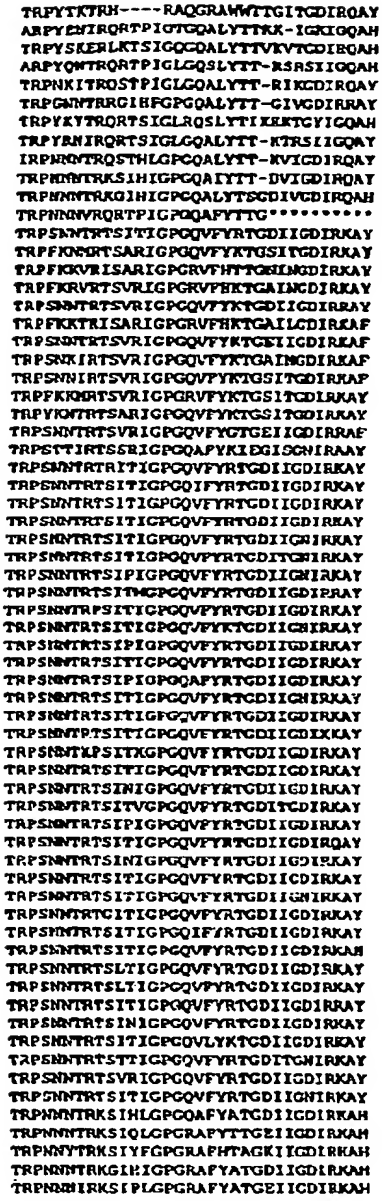
HIV input: 7/10

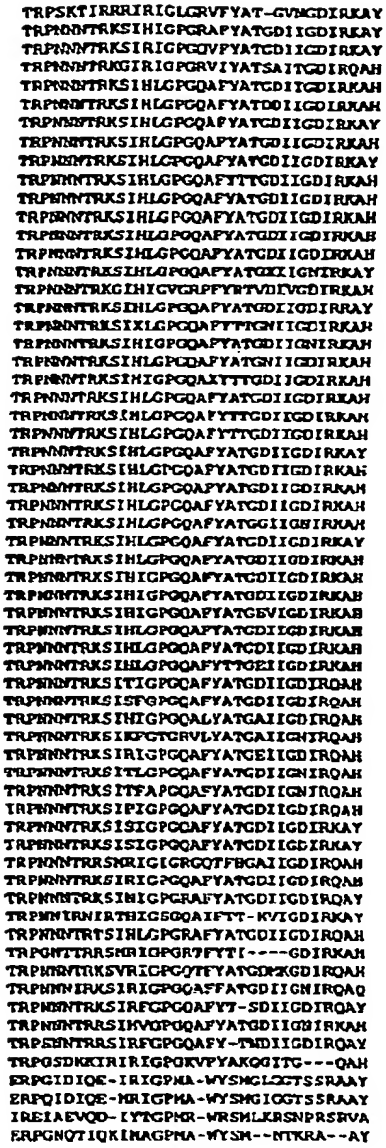


HTV input: 8/ 10



HIV input; 9/10





HIV input: 10/10

補遺C

HIV output: 1/6

```
 0 | 0.00 | & $ A18|Q31|H33 $  & 34019 & 15684.208314 & 0.000000 & \cr
 1 | 0.00 | & $ A18|T21 $  & 33816 & 12352.399254 & 0.000000 & \cr
 2 | 0.01 | & $ A21|D24 $  & 45549 & 17706.407140 & 0.000000 & \cr
 3 | 0.01 | & $ H12|A18 $  & 86025 & 24619.776947 & 0.000000 & \cr
 4 | 0.01 | & $ K13|R17 $  & 88257 & 19028.783592 & 0.000000 & \cr
 5 | 0.01 | & $ I11|R17 $  & 64548 & 27053.952336 & 0.000000 & \cr
 6 | 0.02 | & $ L23|K31 $  & 39387 & 17335.347894 & 0.000000 & \cr
 7 | 0.02 | & $ L13|W19|D24 $  & 20184 & 379.160544 & 0.000000 & \cr
 8 | 0.02 | & $ M13|W15 $  & 23300 & 6673.177086 & 0.000000 & \cr
 9 | 0.02 | & $ N4|K9 $  & 162152 & 74737.922307 & 0.000000 & \cr
10 | 0.03 | & $ N4|K9|H33 $  & 26376 & 5666.716129 & 0.000000 & \cr
11 | 0.03 | & $ Q17|D24 $  & 86991 & 37162.233105 & 0.000000 & \cr
12 | 0.03 | & $ Q31|H33 $  & 233190 & 186078.818611 & 0.000000 & \cr
13 | 0.03 | & $ R12|Q17 $  & 53740 & 10564.956512 & 0.000000 & \cr
14 | 0.04 | & $ R12|T18 $  & 62774 & 28359.197022 & 0.000000 & \cr
15 | 0.04 | & $ R17|A18 $  & 54366 & 27136.625076 & 0.000000 & \cr
16 | 0.04 | & $ R17|E24 $  & 33748 & 10413.255892 & 0.000000 & \cr
17 | 0.04 | & $ R17|Q31 $  & 45065 & 26805.242087 & 0.000000 & \cr
18 | 0.05 | & $ R17|T21 $  & 70301 & 16232.354294 & 0.000000 & \cr
19 | 0.05 | & $ S10|D24 $  & 57772 & 17415.133746 & 0.000000 & \cr
20 | 0.05 | & $ V11|R12 $  & 39546 & 18975.126308 & 0.000000 & \cr
21 | 0.05 | & $ V11|R12|T18 $  & 17620 & 881.251263 & 0.000000 & \cr
22 | 0.06 | & $ K31|Y33 $  & 36346 & 20803.636880 & 0.000002 & \cr
23 | 0.06 | & $ N4|A21 $  & 45441 & 30227.809856 & 0.000003 & \cr
24 | 0.06 | & $ Q17|K31 $  & 25033 & 10875.740384 & 0.000018 & \cr
25 | 0.06 | & $ G10|H12 $  & 20779 & 7151.794446 & 0.000041 & \cr
26 | 0.07 | & $ K9|A21 $  & 40098 & 27695.038620 & 0.000233 & \cr
27 | 0.07 | & $ F19|D24 $  & 29121 & 16875.538795 & 0.000286 & \cr
28 | 0.07 | & $ Q17|A21 $  & 29621 & 18109.021417 & 0.000737 & \cr
29 | 0.07 | & $ H13|E24 $  & 22348 & 10939.327036 & 0.000839 & \cr
30 | 0.08 | & $ N4|K9|I11 $  & 15175 & 4159.316971 & 0.001355 & \cr
31 | 0.08 | & $ S4|T9|T12|V18|R21 $  & 10919 & 1.718549 & 0.001524 & \cr
32 | 0.08 | & $ N4|K9|A21 $  & 11233 & 623.181959 & 0.002105 & \cr
33 | 0.09 | & $ N4|Q31|H33 $  & 21868 & 11328.362393 & 0.002369 & \cr
34 | 0.09 | & $ T19|A21 $  & 44400 & 34516.144368 & 0.004910 & \cr
35 | 0.09 | & $ K9|Q31|H33 $  & 16593 & 6391.723718 & 0.006625 & \cr
36 | 0.09 | & $ W19|Q24 $  & 16738 & 7234.038464 & 0.007331 & \cr
37 | 0.10 | & $ E1|N12 $  & 10844 & 1493.835945 & 0.008575 & \cr
38 | 0.10 | & $ K9|E24 $  & 13847 & 4507.312260 & 0.009408 & \cr
39 | 0.10 | & $ K9|R17 $  & 33735 & 24568.179150 & 0.010326 & \cr
40 | 0.10 | & $ T12|V18 $  & 23076 & 14891.617567 & 0.026159 & \cr
41 | 0.11 | & $ R12|A21 $  & 15497 & 7516.155896 & 0.031231 & \cr
42 | 0.11 | & $ N4|K9|Q31|H33 $  & 8280 & 493.681367 & 0.036905 & \cr
43 | 0.11 | & $ N4|R9|A18 $  & 11655 & 4250.900600 & 0.050618 & \cr
44 | 0.11 | & $ S4|T9|T12|V18|R21|Y33 $  & 7370 & 0.093039 & 0.052023 & \cr
45 | 0.12 | & $ R12|Q17|T18 $  & 7452 & 240.364918 & 0.058991 & \cr
46 | 0.12 | & $ V11|Q17 $  & 14350 & 7329.962834 & 0.068429 & \cr
47 | 0.12 | & $ H12|T21 $  & 23261 & 16324.921094 & 0.072825 & \cr
48 | 0.12 | & $ Q17|Y33 $  & 17288 & 10174.788061 & 0.074203 & \cr
49 | 0.13 | & $ L13|W19 $  & 15536 & 8921.243955 & 0.082437 & \cr
50 | 0.13 | & $ S17|H28 $  & 6529 & 136.997153 & 0.108175 & \cr
51 | 0.13 | & $ N4|K9|Q31 $  & 10228 & 3894.612095 & 0.112708 & \cr
52 | 0.13 | & $ X8|S17 $  & 6573 & 275.512362 & 0.115524 & \cr
53 | 0.14 | & $ R17|Q31|H33 $  & 7265 & 1223.984346 & 0.137215 & \cr
54 | 0.14 | & $ T9|T12|V18|R21 $  & 6003 & 30.417827 & 0.143516 & \cr
55 | 0.14 | & $ N4|K9|A18|H33 $  & 6380 & 549.756091 & 0.157254 & \cr
56 | 0.14 | & $ S10|F19|D24 $  & 6150 & 520.344848 & 0.189437 & \cr
57 | 0.15 | & $ I11|R17|A18 $  & 6555 & 1027.737537 & 0.189642 & \cr
58 | 0.15 | & $ V11|R12|Q17 $  & 5751 & 247.598509 & 0.192378 & \cr
59 | 0.15 | & $ S4|T9|V18|R21 $  & 5514 & 35.313082 & 0.195240 & \cr
60 | 0.15 | & $ S4|T9|T12|V18|R21|K31 $  & 5462 & 0.090571 & 0.197300 & \cr
61 | 0.16 | & $ H12|R17|A18 $  & 5618 & 172.948903 & 0.199184 & \cr
62 | 0.16 | & $ Q9|T11|L19|-23 $  & 5464 & 38.188997 & 0.201464 & \cr
63 | 0.16 | & $ Y4|Q9|T11|-23 $  & 5364 & 35.276055 & 0.213263 & \cr
64 | 0.16 | & $ N4|A18|Q31|H33 $  & 6378 & 1180.344841 & 0.229871 & \cr
65 | 0.17 | & $ L13|N12|R21 $  & 5114 & 15.794611 & 0.243044 & \cr
```

HIV output: 2/6

```
 66 | 0.17 | & $ V13|V15|I19 $  & 5095 & 4.314940 & 0.244059 & \cr
 67 | 0.17 | & $ R12|Q17|D24 $  & 5088 & 122.813489 & 0.261410 & \cr
 68 | 0.18 | & $ S4|T9|V18 $  & 5671 & 868.949180 & 0.285090 & \cr
 69 | 0.18 | & $ G24|S28 $  & 5363 & 579.116112 & 0.297805 & \cr
 70 | 0.18 | & $ S4|T9|R21 $  & 5425 & 650.238691 & 0.289171 & \cr
 71 | 0.18 | & $ K9|I11|R17 $  & 5315 & 590.615207 & 0.296904 & \cr
 72 | 0.19 | & $ V18|K31|Y33 $  & 5524 & 852.751002 & 0.304979 & \cr
 73 | 0.19 | & $ T21|E24 $  & 19192 & 14557.811161 & 0.310756 & \cr
 74 | 0.19 | & $ S4|T9|T12|V18|R21|K31|Y33 $  & 4390 & 0.004904 & 0.350351 & \cr
 75 | 0.19 | & $ S4|T9|V18|R21|Y33 $  & 4341 & 1.910712 & 0.358927 & \cr
 76 | 0.20 | & $ I11|H12|A18 $  & 5225 & 890.707159 & 0.359740 & \cr
 77 | 0.20 | & $ N37|L13 $  & 9314 & 5089.163343 & 0.364791 & \cr
 78 | 0.20 | & $ M1|S12|T20 $  & 4243 & 17.400459 & 0.378694 & \cr
 79 | 0.20 | & $ Y4|T11|-23 $  & 4876 & 710.489341 & 0.388552 & \cr
 80 | 0.21 | & $ H12|A18|K20 $  & 5293 & 1141.301814 & 0.391569 & \cr
 81 | 0.21 | & $ N12|G24|L25 $  & 4169 & 18.987442 & 0.391690 & \cr
 82 | 0.21 | & $ N12|T13 $  & 5365 & 1255.021021 & 0.398803 & \cr
 83 | 0.21 | & $ N4|K9|G23 $  & 9804 & 5726.074196 & 0.404548 & \cr
 84 | 0.22 | & $ P12|L13|W19|Q24 $  & 4070 & 20.998889 & 0.409748 & \cr
 85 | 0.22 | & $ Q12|Y13|T15|G17|V26 $  & 4024 & 0.000255 & 0.414271 & \cr
 86 | 0.22 | & $ S10|F19|A21 $  & 5598 & 1607.067572 & 0.420192 & \cr
 87 | 0.22 | & $ K9|H12 $  & 26788 & 22912.753561 & 0.441631 & \cr
 88 | 0.23 | & $ S10|Q17|D24 $  & 3960 & 93.803024 & 0.443318 & \cr
 89 | 0.23 | & $ Q17|A21|D24 $  & 3949 & 133.058101 & 0.452738 & \cr
 90 | 0.23 | & $ N4|K9|H12 $  & 4239 & 458.472945 & 0.457896 & \cr
 91 | 0.23 | & $ T9|T12|V18|R21|Y33 $  & 3784 & 1.646276 & 0.459063 & \cr
 92 | 0.24 | & $ Y4|Q9|T11 $  & 3402 & 639.401728 & 0.462612 & \cr
 93 | 0.24 | & $ N4|K9|R17 $  & 6239 & 507.820002 & 0.468770 & \cr
 94 | 0.24 | & $ N4|H12|A18 $  & 4450 & 726.477198 & 0.470266 & \cr
 95 | 0.24 | & $ Q9|T11|L19 $  & 4413 & 691.708041 & 0.470653 & \cr
 96 | 0.25 | & $ S4|T9|T12|R21 $  & 3747 & 31.482325 & 0.471755 & \cr
 97 | 0.25 | & $ N12|E30 $  & 4440 & 746.347625 & 0.479764 & \cr
 98 | 0.25 | & $ I1|S2 $  & 3970 & 345.180880 & 0.489218 & \cr
 99 | 0.26 | & $ S4|T12|V18|R21 $  & 3643 & 32.472859 & 0.491921 & \cr
100 | 0.26 | & $ Q9|T11|-23 $  & 4299 & 742.828036 & 0.502461 & \cr
101 | 0.26 | & $ T21|Q31 $  & 16089 & 12621.469597 & 0.519777 & \cr
102 | 0.26 | & $ K9|A18|Q31|N33 $  & 4150 & 697.083962 & 0.520603 & \cr
103 | 0.27 | & $ S4|T9|R21|Y33 $  & 3460 & 35.030273 & 0.528112 & \cr
104 | 0.27 | & $ Y4|Q9|T11|L19|-23 $  & 3425 & 1.024291 & 0.528495 & \cr
105 | 0.27 | & $ S6|K7|T10|L13|N13|K16|G26|Y28 $  & 3409 & 0.000000 & 0.531288 & \cr
106 | 0.27 | & $ S4|T9|V18|R21|R31 $  & 3404 & 1.860057 & 0.532246 & \cr
107 | 0.28 | & $ S17|I19 $  & 4910 & 1510.151983 & 0.533093 & \cr
108 | 0.28 | & $ Y12|H20|R24 $  & 3401 & 29.556849 & 0.538702 & \cr
109 | 0.28 | & $ S4|T9|V18|R21|K31|Y33 $  & 3370 & 0.100690 & 0.539008 & \cr
110 | 0.29 | & $ S10|Q17 $  & 22965 & 16738.120311 & 0.547525 & \cr
111 | 0.29 | & $ A1|-22|S23 $  & 3303 & 7.355264 & 0.553724 & \cr
112 | 0.29 | & $ N13|W15|E31 $  & 3339 & 56.771417 & 0.556389 & \cr
113 | 0.29 | & $ *24|*25|*26|*27|*28|*29|*30|*31|*32|*33 $  & 3269 & 0.000000 & 0.559020 & \r
114 | 0.29 | & $ R17|N33 $  & 31466 & 28229.156188 & 0.565421 & \cr
115 | 0.30 | & $ N13|W15|T28 $  & 3501 & 360.659791 & 0.584679 & \cr
116 | 0.30 | & $ F13|-22|S23 $  & 3121 & 6.681356 & 0.589410 & \cr
117 | 0.30 | & $ R17|A18|T21 $  & 3190 & 85.245047 & 0.592593 & \cr
118 | 0.30 | & $ N4|K9|A18|Q31|N33 $  & 3140 & 55.655691 & 0.594235 & \cr
119 | 0.31 | & $ R17|A18|Q31|N33 $  & 3144 & 101.027645 & 0.604153 & \cr
120 | 0.31 | & $ V1|N23|*24|*25|*26|*27|*28|*29|*30|*31|*32|*33 $  & 3930 & 0.000000 & 0.606
121 | 0.31 | & $ A11|N12 $  & 4517 & 1492.035945 & 0.607916 & \cr
122 | 0.31 | & $ R12|T18|K21 $  & 3150 & 234.485398 & 0.609647 & \cr
123 | 0.32 | & $ S10|G23|D24 $  & 3606 & 599.551395 & 0.611461 & \cr
124 | 0.32 | & $ S1|N13|W15 $  & 3087 & 91.193028 & 0.613590 & \cr
125 | 0.32 | & $ N12|F20|K24 $  & 3202 & 213.735139 & 0.615099 & \cr
126 | 0.32 | & $ N13|W15|E24 $  & 3282 & 306.430052 & 0.617639 & \cr
127 | 0.33 | & $ K9|I11|F19|G23 $  & 4153 & 1180.595212 & 0.618272 & \cr
128 | 0.33 | & $ R2|P3|N5|N6|T7|R8|G10|P15|G16|V20|T23|G23|I25|E26|G27|I29|R30|A32 $  & 3353
129 | 0.33 | & $ N12|A18|Q31 $  & 3759 & 845.163445 & 0.629981 & \cr
130 | 0.34 | & $ K17|D20|-23 $  & 2928 & 25.418797 & 0.632234 & \cr
131 | 0.34 | & $ Y5|K7|R10|R23|W24|T28 $  & 2897 & 0.000008 & 0.633345 & \cr
```

HIV output: 3/6

```
132 | 0.34 | & $ G10|R17 $  & 9506 & 6637.164693 & 0.638967 & \cr
133 | 0.34 | & $ Y4|Q9|-23 $  & 3539 & 699.476594 & 0.644852 & \cr
134 | 0.35 | & $ G12|A22|D23|N24 $  & 2838 & 0.092735 & 0.645134 & \cr
135 | 0.35 | & $ T1|R2|P3|N5|N6|T7|R8|G14|P15|G16|Y20|T22|I25|I26|G27|I29|R30|A32 $  & 3787 &
136 | 0.35 | & $ T1|R2|P3|N5|N6|T7|R8|G14|P15|G16|Y20|T22|G23|I25|I26|G27|I29|R30|A32 $  & 3
137 | 0.35 | & $ N4|H12 $  & 26775 & 23965.157075 & 0.646741 & \cr
138 | 0.36 | & $ V18|*24|*25|*26|*27|*28|*29|*30|*31|*32|*33 $  & 2775 & 0.000000 & 0.657651
139 | 0.36 | & $ A2|R9|-22|S23 $  & 2763 & 0.413224 & 6.660115 & \cr
140 | 0.36 | & $ Y6|Y12|F13|H20|A22|K24 $  & 2761 & 0.000852 & 0.660430 & \cr
141 | 0.36 | & $ V11|A24|S28 $  & 2788 & 31.535138 & 0.661338 & \cr
142 | 0.37 | & $ V20|I22|-24|K25|M29 $  & 2768 & 0.000084 & 0.663009 & \cr
143 | 0.37 | & $ K9|H12|A18 $  & 3267 & 526.343072 & 0.664465 & \cr
144 | 0.37 | & $ T9|T12|V18|R21|K31 $  & 2742 & 1.602621 & 0.664517 & \cr
145 | 0.37 | & $ T8|R9 $  & 3185 & 445.441758 & 0.664693 & \cr
146 | 0.38 | & $ I11|H12|R17 $  & 2909 & 172.776969 & 0.665344 & \cr
147 | 0.38 | & $ T6|K30|K32|N13|K18|K19|R31 $  & 2736 & 0.000000 & 0.665388 & \cr
148 | 0.38 | & $ A24|S28 $  & 3300 & 566.063943 & 0.665797 & \cr
149 | 0.38 | & $ G12|T18|A22|D23|N24 $  & 2692 & 0.005083 & 0.674094 & \cr
150 | 0.39 | & $ P12|W19|Q26 $  & 3054 & 395.340658 & 0.680669 & \cr
151 | 0.39 | & $ A14|H20|N24 $  & 2697 & 47.434702 & 0.682460 & \cr
152 | 0.39 | & $ T9|T12|V18|R21|K31|Y33 $  & 2692 & 0.086760 & 0.685931 & \cr
153 | 0.39 | & $ R12|Q17|A21 $  & 2701 & 79.065229 & 0.687887 & \cr
154 | 0.40 | & $ R17|A18|H33 $  & 3948 & 1325.820319 & 0.688628 & \cr
155 | 0.40 | & $ W15|I19|A24 $  & 2655 & 56.805384 & 0.693552 & \cr
156 | 0.40 | & $ Q12|R13|V20|I22|K24|-26|M29 $  & 2584 & 0.000000 & 0.695328 & \cr
157 | 0.40 | & $ S1|Y4|-6|N10|Y11|S12|S15|V21|K26 $  & 2556 & 0.000000 & 0.701101 & \cr
158 | 0.41 | & $ T38|A21 $  & 6883 & 4332.151205 & 0.701796 & \cr
159 | 0.41 | & $ K17|D20 $  & 2996 & 458.835571 & 0.704460 & \cr
160 | 0.41 | & $ Q17|D24|K31 $  & 2660 & 125.180912 & 0.704916 & \cr
161 | 0.41 | & $ L13|Q15|W19 $  & 2582 & 98.222466 & 0.714812 & \cr
162 | 0.42 | & $ S4|T9|R21|K31|Y33 $  & 2674 & 1.844223 & 0.717056 & \cr
163 | 0.42 | & $ I1|G4|N13|P18|R22|-24|V25 $  & 2445 & 0.000002 & 0.722206 & \cr
164 | 0.42 | & $ S12|F13 $  & 4939 & 2502.178252 & 0.723857 & \cr
165 | 0.43 | & $ L13|Q17|K31 $  & 2663 & 227.867572 & 0.724306 & \cr
166 | 0.43 | & $ K9|R17|H33 $  & 3142 & 710.504406 & 0.724879 & \cr
167 | 0.43 | & $ P12|L13|N25 $  & 2907 & 483.231133 & 0.726360 & \cr
168 | 0.43 | & $ K9|R17|A18 $  & 3012 & 598.308696 & 0.728290 & \cr
169 | 0.44 | & $ S4|T12|R21 $  & 3010 & 597.264141 & 0.728473 & \cr
170 | 0.44 | & $ N4|I11|R17 $  & 3253 & 820.559839 & 0.728529 & \cr
171 | 0.44 | & $ M13|A24|E31 $  & 2426 & 50.435156 & 0.735563 & \cr
172 | 0.44 | & $ L3|A12|T18|V19|D23|A24 $  & 2374 & 0.000104 & 0.735861 & \cr
173 | 0.45 | & $ K9|A21|H33 $  & 3269 & 897.012220 & 0.735243 & \cr
174 | 0.45 | & $ R2|P3|N5|N6|T7|R8|G14|P15|G16|F19|Y20|T22|G23|I25|I26|G27|I29|R30|A32 $  & 3
175 | 0.45 | & $ R10|K11|S12|V25 $  & 2345 & 0.448221 & 0.741446 & \cr
176 | 0.45 | & $ M4|K9|I11|G23 $  & 2883 & 541.944923 & 0.742108 & \cr
177 | 0.46 | & $ R17|A18|Q31 $  & 3304 & 973.769538 & 0.744353 & \cr
178 | 0.46 | & $ Y4|Q9|T11|P13|V19|-23 $  & 2321 & 0.009029 & 0.745895 & \cr
179 | 0.46 | & $ I7|F20|Q33 $  & 2355 & 32.678004 & 0.746775 & \cr
180 | 0.46 | & $ T9|V18|K31|Y33 $  & 2352 & 43.522103 & 0.748251 & \cr
181 | 0.47 | & $ L3|A12|V19|D23|R24 $  & 2307 & 0.001890 & 0.748529 & \cr
182 | 0.47 | & $ G4|M11|P18 $  & 2306 & 12.419975 & 0.751048 & \cr
183 | 0.47 | & $ S4|T12|V18|R21|Y33 $  & 2292 & 1.757250 & 0.151673 & \cr
184 | 0.47 | & $ M13|R17|T21 $  & 2417 & 129.651999 & 0.752215 & \cr
185 | 0.48 | & $ R10|S12|W15|Q24 $  & 2299 & 14.238983 & 0.752700 & \cr
186 | 0.48 | & $ D4|E6|I7|L11|Q12|V13|V20|A22|T24|-25|A26|T29|Q33 $  & 2279 & 0.000003 & 0.75
187 | 0.48 | & $ G10|I24 $  & 2727 & 449.088967 & 0.753366 & \cr
188 | 0.48 | & $ V15|R24|V26|L31 $  & 2272 & 0.404088 & 0.755161 & \cr
189 | 0.49 | & $ V11|R12|Q17|T18 $  & 2281 & 11.909386 & 0.755429 & \cr
190 | 0.49 | & $ L13|W15|Q24|E28 $  & 2270 & 0.994080 & 0.755644 & \cr
191 | 0.49 | & $ T1|R2|P3|N5|N6|T7|R8|G14|P15|G16|F19|Y20|T22|G23|I25|I26|G27|I29|R30|A32 $
192 | 0.49 | & $ R17|T21|E24 $  & 2366 & 123.762808 & 0.760427 & \cr
193 | 0.50 | & $ M13|W15|N25 $  & 2610 & 372.687253 & 0.761541 & \cr
194 | 0.50 | & $ M13|K17|V26 $  & 2455 & 210.504133 & 0.761692 & \cr
195 | 0.50 | & $ M13|Q15|G24 $  & 2335 & 100.386181 & 0.761956 & \cr
196 | 0.51 | & $ F19|G23|D24 $  & 3105 & 885.799386 & 0.764893 & \cr
197 | 0.51 | & $ M11|I15|G18|Q19|T20|F21|N22|A24 $  & 2218 & 0.000000 & 0.765115 & \cr
```

HIV output: 4/6

```
198 | 0.51 | & $ R9|A14|H20|N24 $  & 2214 & 2.664734 & 0.766345 & \cr
199 | 0.51 | & $ S4|T9|V18|K31 $  & 2253 & 45.367685 & 0.767028 & \cr
200 | 0.52 | & $ T4|T11|L19|-23 $  & 2222 & 36.549243 & 0.771106 & \cr
201 | 0.52 | & $ K9|A18|H33 $  & 9877 & 7701.775158 & 0.772980 & \cr
202 | 0.52 | & $ T9|S18|H20 $  & 3246 & 73.652930 & 0.773506 & \cr
203 | 0.52 | & $ G10|S17|I19 $  & 2252 & 85.340274 & 0.774566 & \cr
204 | 0.53 | & $ T12|P13|A14 $  & 2217 & 52.732217 & 0.774983 & \cr
205 | 0.53 | & $ N4|A21|H33 $  & 3446 & 1316.842768 & 0.781367 & \cr
206 | 0.53 | & $ N9|R10|S12|H20|K23|Q24 $  & 2120 & 0.000166 & 0.783023 & \cr
207 | 0.53 | & $ R12|T18|D24 $  & 2373 & 258.082710 & 0.783941 & \cr
208 | 0.54 | & $ T21|H33 $  & 13722 & 11609.774510 & 0.784336 & \cr
209 | 0.54 | & $ T9|V18|R21 $  & 2733 & 627.501151 & 0.785638 & \cr
210 | 0.54 | & $ L13|K17|V19 $  & 2223 & 123.584647 & 0.786733 & \cr
211 | 0.54 | & $ T9|V18|Y33 $  & 2918 & 837.332414 & 0.788304 & \cr
212 | 0.55 | & $ E1|Q4|I5|D6|I7|Q8|E9|-10|M11|M16|A17|-19|M19|S21|M22|I24|G25|G26|T27|S28|S2
213 | 0.55 | & $ Y5|K7|K23|W24|T28 $  & 2075 & 0.000148 & 0.791109 & \cr
214 | 0.55 | & $ S4|W15|I19|A24 $  & 2071 & 3.211998 & 0.792396 & \cr
215 | 0.55 | & $ N4|K9|A18|Q31 $  & 2414 & 350.433693 & 0.793149 & \cr
216 | 0.56 | & $ X12|L13|N26 $  & 2091 & 32.654961 & 0.794078 & \cr
217 | 0.56 | & $ S8|R10|S12|R20|-23|K24 $  & 2056 & 0.001995 & 0.794496 & \cr
218 | 0.56 | & $ S10|A21|D24 $  & 2195 & 141.715389 & 0.794978 & \cr
219 | 0.56 | & $ T5|K6|K7|I8|H10|G24|M26 $  & 2049 & 0.000000 & 0.795739 & \cr
220 | 0.57 | & $ T9|V18|K31 $  & 2861 & 816.350692 & 0.796510 & \cr
221 | 0.57 | & $ I20|A22|T23|K24 $  & 2040 & 0.135518 & 0.797358 & \cr
222 | 0.57 | & $ Y3|A4|-21|N23 $  & 2039 & 0.001752 & 0.797511 & \cr
223 | 0.57 | & $ G4|N11|D23|G24 $  & 2039 & 0.335758 & 0.797570 & \cr
224 | 0.58 | & $ I11|E24 $  & 4624 & 2601.997572 & 0.798748 & \cr
225 | 0.58 | & $ G10|G17|G24 $  & 2157 & 138.303116 & 0.801095 & \cr
226 | 0.58 | & $ Y6|S8|R10|A15|R16|K22|K24 $  & 2011 & 0.000000 & 0.802468 & \cr
227 | 0.59 | & $ S4|T9|R21|K31 $  & 2063 & 30.105157 & 0.802810 & \cr
228 | 0.59 | & $ D4|E6|I7|R9|L11|Q12|V13|V20|A22|T24|-25|A26|T29|Q33 $  & 1999 & 0.000000 &
229 | 0.59 | & $ Q9|T11|L19|-23|K24 $  & 1990 & 1.569195 & 0.806400 & \cr
230 | 0.59 | & $ S8|P12|K24 $  & 2000 & 16.301963 & 0.807225 & \cr
231 | 0.60 | & $ S4|T9|T12|R21|Y33 $  & 1985 & 1.703737 & 0.807295 & \cr
232 | 0.60 | & $ R10|V12|V19|Q24|R31 $  & 1982 & 0.041977 & 0.807529 & \cr
233 | 0.60 | & $ T4|Q9|F20|K23|G24 $  & 1979 & 0.004973 & 0.808044 & \cr
234 | 0.60 | & $ L11|S12|L13|V26 $  & 1972 & 4.533048 & 0.810047 & \cr
235 | 0.61 | & $ T5|K6|K7|I8|N9|H10|G24|M26 $  & 1967 & 0.000000 & 0.810130 & \cr
236 | 0.61 | & $ S6|K7|T10|L11|K16|G26|Y28 $  & 1956 & 0.000000 & 0.812033 & \cr
237 | 0.61 | & $ T9|V18|R21|Y33 $  & 1983 & 33.839576 & 0.811214 & \cr
238 | 0.61 | & $ R2|P3|N5|N6|T7|R8|G14|P15|G16|Y20|T22|G23|I25|Z26|G27|D28|Z29|R30|A32 $  &
239 | 0.62 | & $ E19|A21|D24 $  & 2084 & 139.085768 & 0.813940 & \cr
240 | 0.62 | & $ L11|N13|W15 $  & 1949 & 9.905173 & 0.814948 & \cr
241 | 0.62 | & $ Q9|T11|Q12|L19|F20|K22|T23|R24 $  & 1933 & 0.000000 & 0.815996 & \cr
242 | 0.62 | & $ F19|A21|G23 $  & 4316 & 2404.525279 & 0.816257 & \cr
243 | 0.63 | & $ H12|R17|Z24 $  & 2006 & 91.386294 & 0.819181 & \cr
244 | 0.63 | & $ L13|W15|V26 $  & 2169 & 237.129335 & 0.819611 & \cr
245 | 0.63 | & $ N12|M19|-23|N24 $  & 1909 & 7.808758 & 0.821430 & \cr
246 | 0.63 | & $ T1|R2|P3|N5|N6|T7|R8|G14|P15|G16|Y20|T22|I25|G27|I29|R30|A32 $  & 4991 & 3
247 | 0.64 | & $ T21|Q24 $  & 7773 & 5882.829451 & 0.823380 & \cr
248 | 0.64 | & $ G4|V11|R12 $  & 2497 & 605.660868 & 0.823510 & \cr
249 | 0.64 | & $ Q17|R31|Y33 $  & 2149 & 263.756833 & 0.824134 & \cr
250 | 0.64 | & $ K25|K26 $  & 2095 & 217.906699 & 0.825341 & \cr
251 | 0.65 | & $ T21|Q31|H33 $  & 2236 & 361.582066 & 0.825942 & \cr
252 | 0.65 | & $ T12|V18|R21 $  & 3446 & 576.530816 & 0.826794 & \cr
253 | 0.65 | & $ Y4|Q9|T11|L19|-23|N24 $  & 1865 & 0.047981 & 0.826881 & \cr
254 | 0.65 | & $ H12|Q31|H33 $  & 2922 & 1055.147497 & 0.827268 & \cr
255 | 0.66 | & $ R9|M11|I15|G18|Q19|T20|F21|N22|A24 $  & 1865 & 0.000000 & 0.827546 & \cr
256 | 0.66 | & $ V1|T18|N23|*24|*25|*26|*27|*28|*29|*30|*31|*32|*33 $  & 1859 & 0.000000 &
257 | 0.66 | & $ G4|L13|W15|A24 $  & 1866 & 7.095250 & 0.828568 & \cr
258 | 0.66 | & $ P12|T21 $  & 11256 & 9405.119546 & 0.829912 & \cr
259 | 0.67 | & $ T9|Y31|Y33 $  & 2569 & 823.122139 & 0.830747 & \cr
260 | 0.67 | & $ K7|A14|A24|V32 $  & 1844 & 0.130705 & 0.831083 & \cr
261 | 0.67 | & $ Q9|T11|I19|L22|F23|K24|V25|V26 $  & 1841 & 0.000000 & 0.831561 & \cr
262 | 0.68 | & $ R9|K17|G24 $  & 2157 & 318.307130 & 0.831945 & \cr
263 | 0.68 | & $ A14|G24 $  & 5018 & 3183.960671 & 0.832719 & \cr
```

HIV output 5/6

```
268 | 0.68 | & $ S8|R10|S12|V20|A22|R23 $  & 1834 & 0.000248 & 0.832736 & \cr
265 | 0.68 | & $ L11|S12|V26 $  & 1919 & 85.184287 & 0.832757 & \cr
266 | 0.69 | & $ R2|P3|N5|R6|T7|R8|S10|G14|P15|G16|F19|Y20|T22|G23|I25|I26|G27|I29|R30|A32 $
267 | 0.69 | & $ R12|T18|R31 $  & 2316 & 510.426023 & 0.834125 & \cr
268 | 0.69 | & $ V11|R12|D24 $  & 2089 & 265.729908 & 0.834506 & \cr
269 | 0.69 | & $ H12|A18|T21 $  & 1899 & 83.244349 & 0.835749 & \cr
270 | 0.70 | & $ R12|D24 $  & 12816 & 11016.726531 & 0.838461 & \cr
271 | 0.70 | & $ D5|Q24|N28 $  & 1856 & 57.570035 & 0.838602 & \cr
272 | 0.70 | & $ V11|A21 $  & 6489 & 4695.336767 & 0.839384 & \cr
273 | 0.70 | & $ T9|T12|R21 $  & 2384 & 550.736361 & 0.[illegible]758 & \cr
274 | 0.71 | & $ G10|L13|V19|Q24 $  & 1805 & 21.432425 & 0.841035 & \cr
275 | 0.71 | & $ S4|K10 $  & 2670 & 889.423177 & 0.841523 & \cr
276 | 0.71 | & $ Q9|L19|-23|R24|V25 $  & 1781 & 0.559631 & 0.841545 & \cr
277 | 0.71 | & $ H12|V19|N28 $  & 1920 & 140.804894 & 0.841748 & \cr
278 | 0.72 | & $ H4|R5|T12 $  & 1843 & 63.837063 & 0.841754 & \cr
279 | 0.72 | & $ R10|S12|S19|Q24 $  & 1775 & 2.690479 & 0.842869 & \cr
280 | 0.72 | & $ K13|R17|T18 $  & 2153 & 386.143038 & 0.843755 & \cr
281 | 0.72 | & $ R17|T21|Q31 $  & 1850 & 51.352915 & 0.845085 & \cr
282 | 0.73 | & $ S8|K24 $  & 2047 & 293.960536 & 0.845991 & \cr
283 | 0.73 | & $ Y6|Q9|R10|T11|L19|-23|R24 $  & 1749 & 0.003580 & 0.846684 & \cr
284 | 0.73 | & $ I1|E3|I4|A5|E6|V7|Q8|D9|-10|Y12|T13|K16|-18|W19|R20|S21|N22|L23|K26|N29|S26
285 | 0.73 | & $ D5|Q24 $  & 2759 & 1018.937653 & 0.848081 & \cr
286 | 0.74 | & $ T1|R2|P3|N5|N6|T7|R8|G14|P15|G16|Y20|T22|I25|I26|G27|D28|I29|R30|A32 $  & 1:
287 | 0.74 | & $ K9|H12|R17 $  & 1894 & 166.406797 & 0.850079 & \cr
288 | 0.74 | & $ V11|Q17|D24 $  & 1812 & 93.103584 & 0.851467 & \cr
289 | 0.74 | & $ A1|H11 $  & 2729 & 1013.253017 & 0.851968 & \cr
290 | 0.75 | & $ H12|A18|Q31|H33 $  & 1795 & 85.511085 & 0.852963 & \cr
291 | 0.75 | & $ L19|S22|V26 $  & 1757 & 49.527891 & 0.853283 & \cr
292 | 0.75 | & $ R9|N12|M13 $  & 2146 & 444.904483 & 0.854292 & \cr
293 | 0.76 | & $ Q6|M13|W15|W20|N22 $  & 1695 & 0.065297 & 0.855256 & \cr
294 | 0.76 | & $ Y6|T12|L19 $  & 2355 & 661.700054 & 0.855514 & \cr
295 | 0.76 | & $ I19|-23|V25 $  & 1827 & 134.704350 & 0.855682 & \cr
296 | 0.76 | & $ T9|V18|R21|K33|Y33 $  & 1692 & 1.781938 & 0.856009 & \cr
297 | 0.77 | & $ S15|-21|-22|-24 $  & 1664 & 0.060460 & 0.860125 & \cr
298 | 0.77 | & $ K12|N28 $  & 2272 & 614.028472 & 0.861054 & \cr
299 | 0.77 | & $ A1|M11|T18 $  & 1713 & 55.464572 & 0.861122 & \cr
300 | 0.77 | & $ V20|I22|-24|R25|N28|H29 $  & 1657 & 0.000005 & 0.861205 & \cr
301 | 0.78 | & $ H9|N28 $  & 3094 & 448.609779 & 0.863034 & \cr
302 | 0.78 | & $ A21|Q31|H33 $  & 3220 & 1574.859557 & 0.863042 & \cr
303 | 0.78 | & $ Q12|R13|V20|I22|K24|-26|N28|H29 $  & 1645 & 0.000000 & 0.863064 & \cr
304 | 0.78 | & $ L3|T9 $  & 3676 & 2021.126046 & 0.863083 & \cr
305 | 0.79 | & $ D24|K31 $  & 12967 & 11324.565776 & 0.863860 & \cr
306 | 0.79 | & $ L13|K31|Y33 $  & 2465 & 827.871734 & 0.866278 & \cr
307 | 0.79 | & $ L19|T21|-23 $  & 1948 & 313.990923 & 0.866436 & \cr
308 | 0.79 | & $ G4|E9|R10|S15|I20|N28|Q24 $  & 1633 & 0.000021 & 0.864913 & \cr
309 | 0.80 | & $ G6|W15|A24 $  & 1765 & 133.549343 & 0.865121 & \cr
310 | 0.80 | & $ T1|R2|P3|N5|N6|T7|R8|G14|P15|G16|Y20|T22|G23|I25|I26|G27|D28|I29|R30|A32 $
311 | 0.80 | & $ N6|R17|A18 $  & 2464 & 833.718080 & 0.865331 & \cr
312 | 0.80 | & $ Y6|T8|K8|H9|-10|-11|-12|-13|R14|A15|Q16|G17|R18|A19|W20|W21|T23|G24|T26 $
313 | 0.81 | & $ H13|S17|T19 $  & 1697 & 69.061643 & 0.865691 & \cr
314 | 0.81 | & $ E13|R17|Q31 $  & 2857 & 1234.208538 & 0.866479 & \cr
315 | 0.81 | & $ S1|V11|V25 $  & 1725 & 114.913060 & 0.868418 & \cr
316 | 0.81 | & $ Q9|L19|-23 $  & 2367 & 758.375850 & 0.868641 & \cr
317 | 0.82 | & $ A12|V19|R24 $  & 1625 & 17.158947 & 0.869764 & \cr
318 | 0.82 | & $ K8|P9|K13|K31 $  & 1606 & 0.000107 & 0.869840 & \cr
319 | 0.82 | & $ P12|A14|S17|W19|Y20 $  & 1665 & 0.078275 & 0.869203 & \cr
320 | 0.82 | & $ E1|K8|T12|D20|V22|Q24 $  & 1601 & 0.000007 & 0.869647 & \cr
321 | 0.83 | & $ T11|L19|-23 $  & 2366 & 770.136365 & 0.870576 & \cr
322 | 0.83 | & $ I1|R9|I12 $  & 3015 & 19.394937 & 0.870616 & \cr
323 | 0.83 | & $ H3|R12 $  & 2021 & 425.578976 & 0.870643 & \cr
324 | 0.84 | & $ I1|I11 $  & 1939 & 345.480880 & 0.870930 & \cr
325 | 0.84 | & $ R9|S15|-21|-21|-23|-24 $  & 1592 & 0.000188 & 0.871160 & \cr
326 | 0.84 | & $ A12|T18|V19|R24 $  & 1583 & 0.942061 & 0.872657 & \cr
327 | 0.84 | & $ W15|Q24|E28 $  & 1596 & 18.677870 & 0.873368 & \cr
328 | 0.85 | & $ Y12|V19|Q24|R31 $  & 1578 & 0.826255 & 0.873390 & \cr
329 | 0.85 | & $ E1|G4|I5|D6|I7|Q8|E9|-10|W16|A17|-19|W19|S21|N22|L24|G25|G26|T27|S28|S29|A3
```

HIV output: 6/6

```
330 | 0.85 | & $ G8|L13|A14|G18|H20 $  & 1575 & 0.000857 & 0.873716 & \cr
331 | 0.85 | & $ M11|R12|T18 $  & 1704 & 336.819937 & 0.874586 & \cr
332 | 0.86 | & $ Y4|I6|Q9|T11|Y19|I23|S24|V25 $  & 1569 & 0.000000 & 0.874613 & \cr
333 | 0.86 | & $ S4|K7|T9|A14|A24|V32 $  & 1568 & 0.000391 & 0.874763 & \cr
334 | 0.86 | & $ G10|V19|R24|V26|L31 $  & 1567 & 0.022878 & 0.874915 & \cr
335 | 0.86 | & $ A18|T21|H33 $  & 1950 & 386.091755 & 0.875371 & \cr
336 | 0.87 | & $ T9|V18|R21|K31 $  & 1595 & 32.945102 & 0.875649 & \cr
337 | 0.87 | & $ N12|H28|E31 $  & 1644 & 84.308230 & 0.876001 & \cr
338 | 0.87 | & $ H4|K9|F19|G23 $  & 2413 & 853.445657 & 0.876021 & \cr
339 | 0.87 | & $ S15|-21|-22|-23|-24 $  & 1550 & 0.003156 & 0.877439 & \cr
340 | 0.88 | & $ V15|P18|R21|Y23|V26 $  & 1543 & 0.000396 & 0.878473 & \cr
341 | 0.88 | & $ V11|R12|A21 $  & 1677 & 139.056915 & 0.879218 & \cr
342 | 0.88 | & $ S4|K11|Y33 $  & 2439 & 891.874986 & 0.879339 & \cr
343 | 0.88 | & $ Y4|Q9|T11|L19 $  & 1566 & 32.897928 & 0.879930 & \cr
344 | 0.89 | & $ A14|S17|W19|P20 $  & 1534 & 1.410979 & 0.880025 & \cr
345 | 0.89 | & $ G4|A9|F20|T26 $  & 1540 & 12.834455 & 0.880809 & \cr
346 | 0.89 | & $ Y6|X10|X12|X18|X19|X31 $  & 1525 & 0.000000 & 0.881117 & \cr
347 | 0.89 | & $ L1|S4|K13|W15 $  & 1525 & 0.559824 & 0.881199 & \cr
348 | 0.90 | & $ M4|K9|A21|H33 $  & 1568 & 48.227041 & 0.881881 & \cr
349 | 0.90 | & $ T9|V11|R22 $  & 1769 & 253.858233 & 0.882556 & \cr
350 | 0.90 | & $ Y6|G8|R10|L11|S12|V20|R23|K24 $  & 1515 & 0.000000 & 0.882576 & \cr
351 | 0.90 | & $ X4|H31 $  & 1926 & 418.267274 & 0.883632 & \cr
352 | 0.91 | & $ P12|D23|N24 $  & 1623 & 115.955006 & 0.883732 & \cr
353 | 0.91 | & $ Q9|K23 $  & 4487 & 2984.952760 & 0.884744 & \cr
354 | 0.91 | & $ G4|R9|M13|W15 $  & 1511 & 14.154263 & 0.885206 & \cr
355 | 0.91 | & $ R2|P3|H5|N6|T7|R8|I13|G14|P15|G16|F19|Y20|T22|G23|I25|I26|G27|I29|R30|A32 $
356 | 0.92 | & $ V13|W15|L19 $  & 1573 & 83.913345 & 0.886323 & \cr
357 | 0.92 | & $ P12|S30 $  & 2786 & 1259.604637 & 0.886566 & \cr
358 | 0.92 | & $ V1|R12|T18|N23|*24|*25|*26|*27|*28|*29|*30|*31|*32|*33 $  & 1487 & 0.000000
359 | 0.93 | & $ Q17|D24|K33 $  & 1608 & 121.315232 & 0.886668 & \cr
360 | 0.93 | & $ E9|R12|T18 $  & 1616 & 133.606703 & 0.887568 & \cr
361 | 0.93 | & $ G4|R12|T18 $  & 2078 & 597.832556 & 0.887601 & \cr
362 | 0.93 | & $ H4|P12 $  & 2777 & 1298.604637 & 0.887855 & \cr
363 | 0.94 | & $ T1|R2|P3|H5|N6|T7|R8|G14|G16|Y20|T22|G23|I25|I26|G27|I29|R30|A32 $  & 1925 &
364 | 0.94 | & $ S4|T9|N12|V18|R21 $  & 1474 & 1.158008 & 0.888647 & \cr
365 | 0.94 | & $ W19|K24|T26 $  & 1724 & 252.469231 & 0.888834 & \cr
366 | 0.94 | & $ A1|E9 $  & 2089 & 630.489963 & 0.890681 & \cr
367 | 0.95 | & $ A1|G4|F20|A24 $  & 1455 & 2.044033 & 0.891465 & \cr
368 | 0.95 | & $ T9|C12|-22|G24 $  & 1450 & 0.214607 & 0.891912 & \cr
369 | 0.95 | & $ Y4|Q9|T11|Y19|W20|-23|N28 $  & 1447 & 0.000061 & 0.892304 & \cr
370 | 0.95 | & $ G10|M13|A14|E31 $  & 1446 & 2.855158 & 0.892845 & \cr
371 | 0.96 | & $ S10|D24|I26 $  & 2229 & 789.361165 & 0.893336 & \cr
372 | 0.96 | & $ G4|M13|W15 $  & 1691 & 252.133281 & 0.893446 & \cr
373 | 0.96 | & $ N12|E31 $  & 2929 & 1492.835945 & 0.893827 & \cr
374 | 0.96 | & $ T12|F13|A14|N28 $  & 1436 & 2.983773 & 0.894262 & \cr
375 | 0.97 | & $ S4|T12|V18|R21|K33 $  & 1436 & 1.710658 & 0.894363 & \cr
376 | 0.97 | & $ S8|G10|X24 $  & 1444 & 16.637502 & 0.895049 & \cr
377 | 0.97 | & $ Q9|T11|L19|-23|K24|R31 $  & 1427 & 0.051495 & 0.895156 & \cr
378 | 0.97 | & $ M13|S17|N28 $  & 1661 & 239.347729 & 0.895842 & \cr
379 | 0.98 | & $ R10|Y12|V19|E23|Q24 $  & 1420 & 0.021913 & 0.896074 & \cr
380 | 0.98 | & $ R12|I15 $  & 2328 & 909.283714 & 0.896119 & \cr
381 | 0.98 | & $ G10|I26|A32|T33|K24 $  & 1415 & 0.007677 & 0.896763 & \cr
382 | 0.98 | & $ Q9|V18|K23 $  & 1572 & 162.169493 & 0.897472 & \cr
383 | 0.99 | & $ R17|T21|H33 $  & 1486 & 82.159457 & 0.898299 & \cr
384 | 0.99 | & $ T9|S18|K30 $  & 1425 & 25.486414 & 0.898892 & \cr
385 | 0.99 | & $ G8|A14|G18|H20 $  & 1399 & 0.016110 & 0.898964 & \cr
386 | 0.99 | & $ Y12|S15|H20|I22|E23|K24 $  & 1393 & 0.000820 & 0.899702 & \cr
387 | 1.00 | & $ I1|Y4|-22|S23|K24 $  & 1393 & 0.040715 & 0.899788 & \cr
```

補遺D

File probsort.pl: 1/1

```
$fm = $ARGV[0];

open (IN, $fm);
@prob = <IN>;
chop @prob;
close (IN);

@prob = grep (/cr/, @prob);

open (TEMP, "> probsort.temp");

foreach (@prob)
{
   print TEMP "$_\n";
}

close (TEMP);

# exit;

$fm = $fm . ".prob";

# print "fm: $fm\n";

`sort  -o prob.tmp -n012345678$0. +9 probsort.temp`;
`rm probsort.temp`;

open (IN, "prob.tmp");
@m2 = <IN>;
chop @m2;
close (IN);

`rm prob.tmp`;

open (TEMP, "> $fm");

$total = scalar @m2;
$i = 0;
foreach (@m2)
{
   printf TEMP "%3d | %.2f | %s\n", $i, (($i / $total)), $_;
   $i++;
}
```

補遺E

A[GAP43 RATGAP43] D[nAChRd RNZCRD1] A[PTN RATHBGAM] D[Ins1 RNINS1] B[cjun RNRJG9] A[CCO1 RATMTCYTOC] A[DD63.2 (I I)]	A[ODC RATODC] D[nAChRe RNACRE] B[FGFRRATFGFR1] A[cyclin A RATPCNA] A[TCP (I I)] A[CCO2 RATMTCYTOC]	B[nAChRa4 RATNARAA] A[CNTFR S54212] B[TGFR RATTGFBIIR] A[H2AZ RATHIS2AZ] F[actin RNAC01] A[SC1 RNU19135]	0.00000
A[GAP43 RATGAP43] B[nAChRa4 RATNARAA] A[CCO2 RATMTCYTOC]	A[GAD65 RATGAD65] B[FGFR RATFGFR1]	A[GRg2 (#)] B[cjun RNRJG9]	0.00000
A[GAP43 RATGAP43] C[mGluR2 RATMGLURB] B[FGFR RATFGFR1] B[SOD RNSODR]	F[NFM RATNFM] B[nAChRa4 RATNARAA] A[IGF I RATIGFIA] A[CCO2 RATMTCYTOC]	A[G67I80/86 RATGAD67] B[nAChRa5 RATNACHRR] B[cjun RNRJG9]	0.00000
B[NMDA2D RNU08260] B[EGF RATEPGF]	D[nAChRe RNACRE] B[TGFR RATTGFBIIR]	C[mAChR4 RATACHRMD]	0.00000
B[G67I80/86 RATGAD67]	A[SOD RNSODR]	B[SC7 RNU19141]	0.00000
A[MAP2 RATMAP2] A[synaptophysin RNSYN] B[ChAT (*)] A[GRa2 (|)] A[GRb3 RATGARB3] C[mGluR8 MMU17252] D[nAChRa6 RATNARA6S] A[trkB RATTRKB1] A[PTN RATHBGAM] B[IGF II RATGFI2] A[H2AZ RATHIS2AZ] A[TCP (I I)] A[CCO2 RATMTCYTOC] A[DD63.2 (I I)]	A[GAP43 RATGAP43] A[neno RATENONS] A[ODC RATODC] A[GRa3 RNGABAA] A[GRg3 RATGABAA] B[NMDA2B RATNMDA2B] D[nAChRd RNZCRD1] A[CNTFR S54212] B[FGFR RATFGFR1] A[IP3R2 RNITPR2R] B[cjun RNRJG9] F[actin RNAC01] A[SC1 RNU19135]	B[L1 S55536] F[GAT1 RATGABAT] B[NOS RRBNOS] A[GRa5 (#)] B[mGluR3 RATMGLURC] B[nAChRa4 RATNARAA] B[5HT1b RAT5HT1BR] A[MK2 MUSMK] D[Ins1 RNINS1] A[cyclin A RATPCNA] B[Brm (I I)] A[CCO1 RATMTCYTOC] D[SC6 RNU19140]	0.00000
D[GFAP RNU03700] A[NT3 RATHDNFNT] C[PDGFb RNPDGFBCP] C[cfos RNCFOSR]	D[GRb2 RATGARB2] B[CNTF RNCNTF] B[PDGFR RNPDGFRBE]	D[NMDA2C RATNMCA2C] D[bFGF RNFGFT] A[cyclin B RATCYCLNB]	0.00000
F[cellubrevin s63830] B[InsR RATINSAB]	D[G67I86 RATGAD67]	B[IGF I RATIGFIA]	0.00000
A[GAD67 RATGAD67] B[5HT2 RATSR5HT2]	C[mGluR6 RATMGLUR6.] B[Ins2 RNINS2]	C[mAChR3 RATACHRMB]	0.00000
F[cellubrevin s63830] B[InsR RATINSAB]	D[mGluR6 RATMGLUR6.] A[SC2 RNU19136]	D[5HT3 MOUSE5HT3]	0.00000
A[nestin RATNESTIN] B[CNTF RNCNTF]	B[TH RATTOHA] B[EGF RATEPGF]	C[mAChR4 RATACHRMD]	0.00000

A[nestin RATNESTIN] A[MK2 MUSMK]	B[TH RATTOHA] B[IGF II RATGFI2]	C[NGF RNNGFB] B[Brm (I I)]	0.00000
A[ODC RATODC] C[NGF RNNGFB] A[MK2 MUSMK] D[Ins1 RNINS1] A[H2AZ RATHIS2AZ] F[actin RNAC01] A[DD63.2 (I I)]	D[nAChRd RNZCRD1] D[trk RATTRKPREC] A[PTN RATHBGAM] B[IGF II RATGFI2] B[Brm (I I)] A[CCO1 RATMTCYTOC]	D[nAChRe RNACRE] A[CNTFR S54212] B[TGFR RATTGFBIIR] A[cyclin A RATPCNA] A[TCP (I I)] A[SC1 RNU19135]	0.00000
A[GAP43 RATGAP43] B[nAChRa4 RATNARAA] B[FGFR RATFGFR1]	F[NFM RATNFM] B[nAChRa5 RATNACHRR] B[cjun RNRJG9]	C[mGluR2 RATMGLURB] B[trkC RATTRKCN3] A[CCO2 RATMTCYTOC]	0.00000
B[TH RATTOHA] B[Brm (I I)]	A[MK2 MUSMK]	B[IGF II RATGFI2]	0.00000
D[mGluR1 RATGPCGR] A[EGFR RATEGFR]	D[mGluR4 RATMGLUR4B] A[IGFR1 RATIGFI]	D[nAChRa2 RATNNAR] A[IGFR2 MMU04710]	0.00000
F[NFL RATNFL] D[nAChRa2 RATNNAR] A[SC2 RNU19136]	D[mGluR4 RATMGLUR4B] D[5HT3 MOUSE5HT3]	D[mGluR6 RATMGLUR6.] A[IGFR1 RATIGFI]	0.00000
D[MOG RATMOG] D[mGluR4 RATMGLUR4B] A[IGFR2 MMU04710]	B[GRa1 (#)] D[nAChRa2 RATNNAR] C[IP3R3 RATIP3R3X]	D[mGluR1 RATGPCGR] A[EGFR RATEGFR]	0.00000
A[GAP43 RATGAP43] B[nAChRa5 RATNACHRR] B[cjun RNRJG9]	F[NFM RATNFM] B[FGFR RATFGFR1] A[CCO2 RATMTCYTOC]	B[nAChRa4 RATNARAA] A[IGF I RATIGFIA]	0.00000
A[cellubrevin s63830] A[CRAF RATRAFA]	A[GRb1 RATGARB1] B[IP3R1 RATI145TR]	A[IGF I RATIGFIA]	0.00000
B[keratin RNKER19] B[CNTF RNCNTF]	A[cellubrevin s63830] A[IGF I RATIGFIA]	B[TH RATTOHA] A[InsR RATINSAB]	0.00000

【図１】

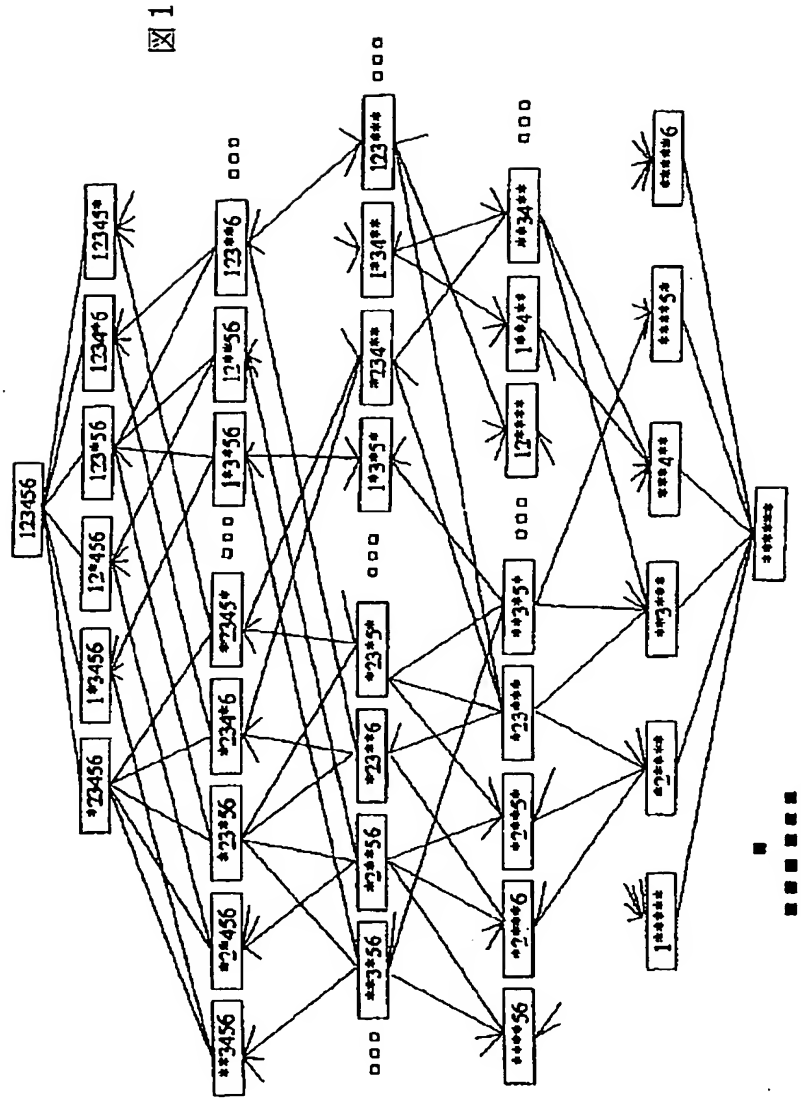
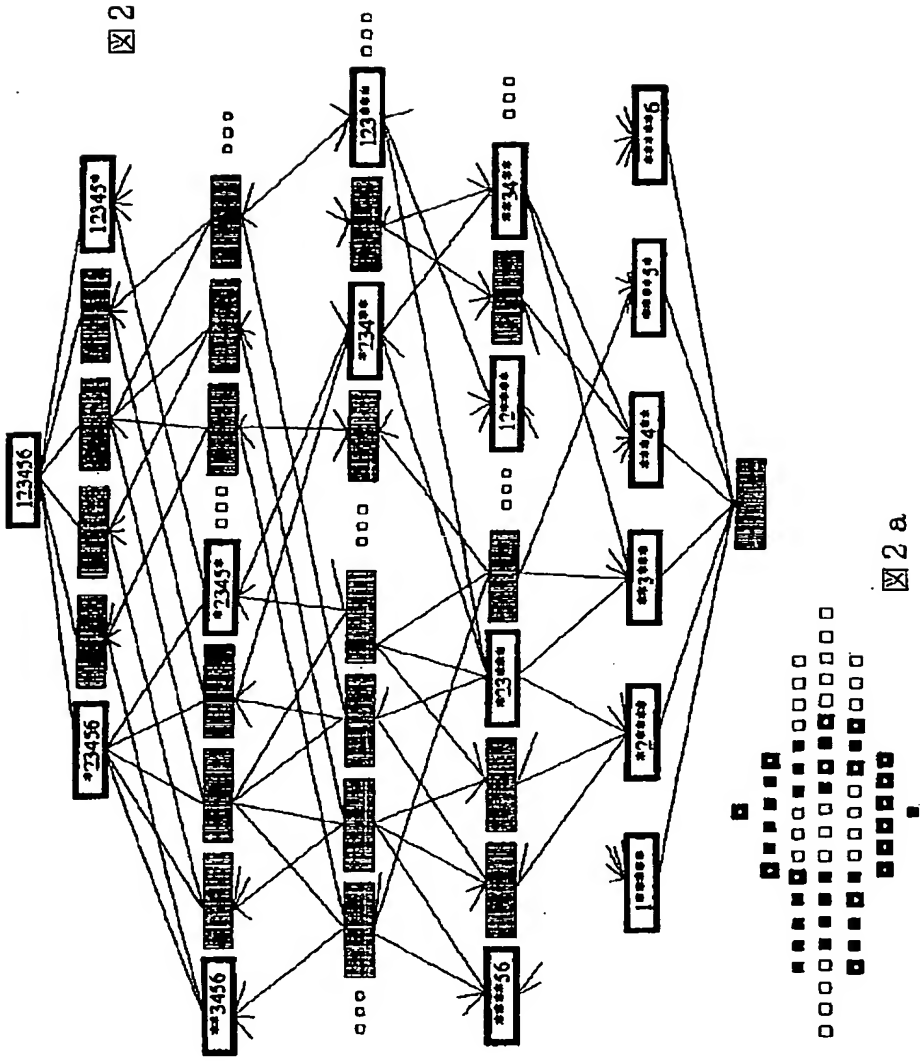

123456
12345*
1234*6
123*56
12*456
1*3456
*23456
12****
1*34**
1**4**
1***5*
1****6
1*****
123***
12*3**
12*45*
12*456
1*3***
1*3*5*
1*3*56
1**45*
1**456
1***56
*234**
*23***
*23*5*
*23*56
*2**5*
*2**56
*2***6
34
3*
**3*5*
**3*56
**345*
**3456
***45*
***456
****56
****5*
****56
***4**
3*
*2****
1*****
******

図１ａ

【図２】



【図３】

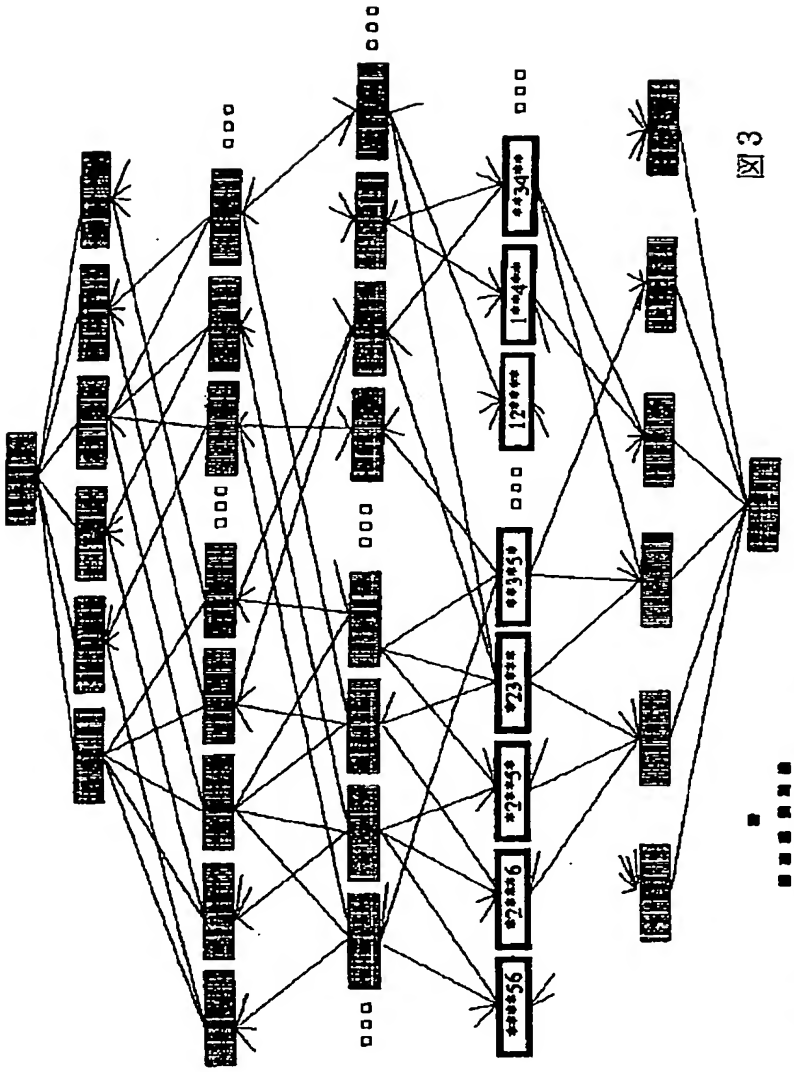


図３

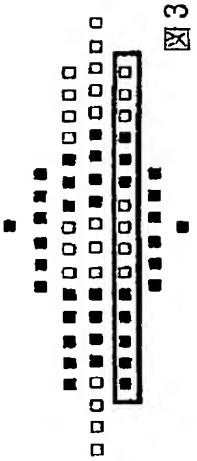
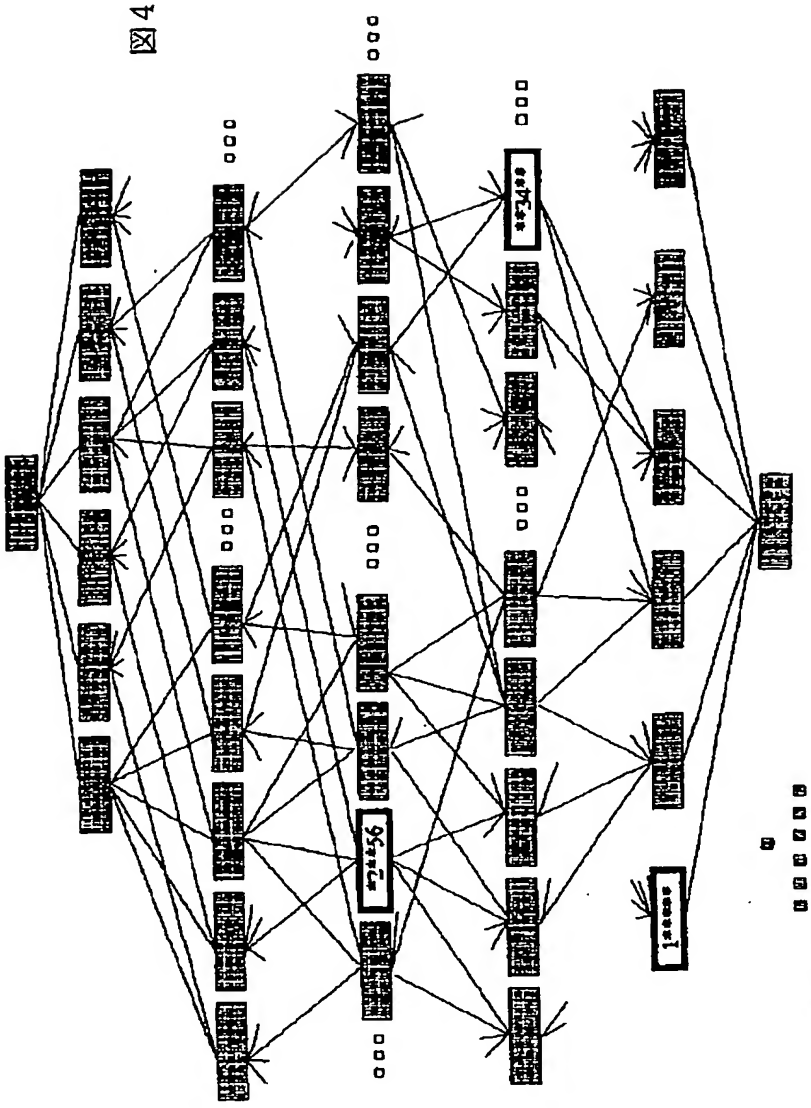


図３ａ

【図４】



【図５】

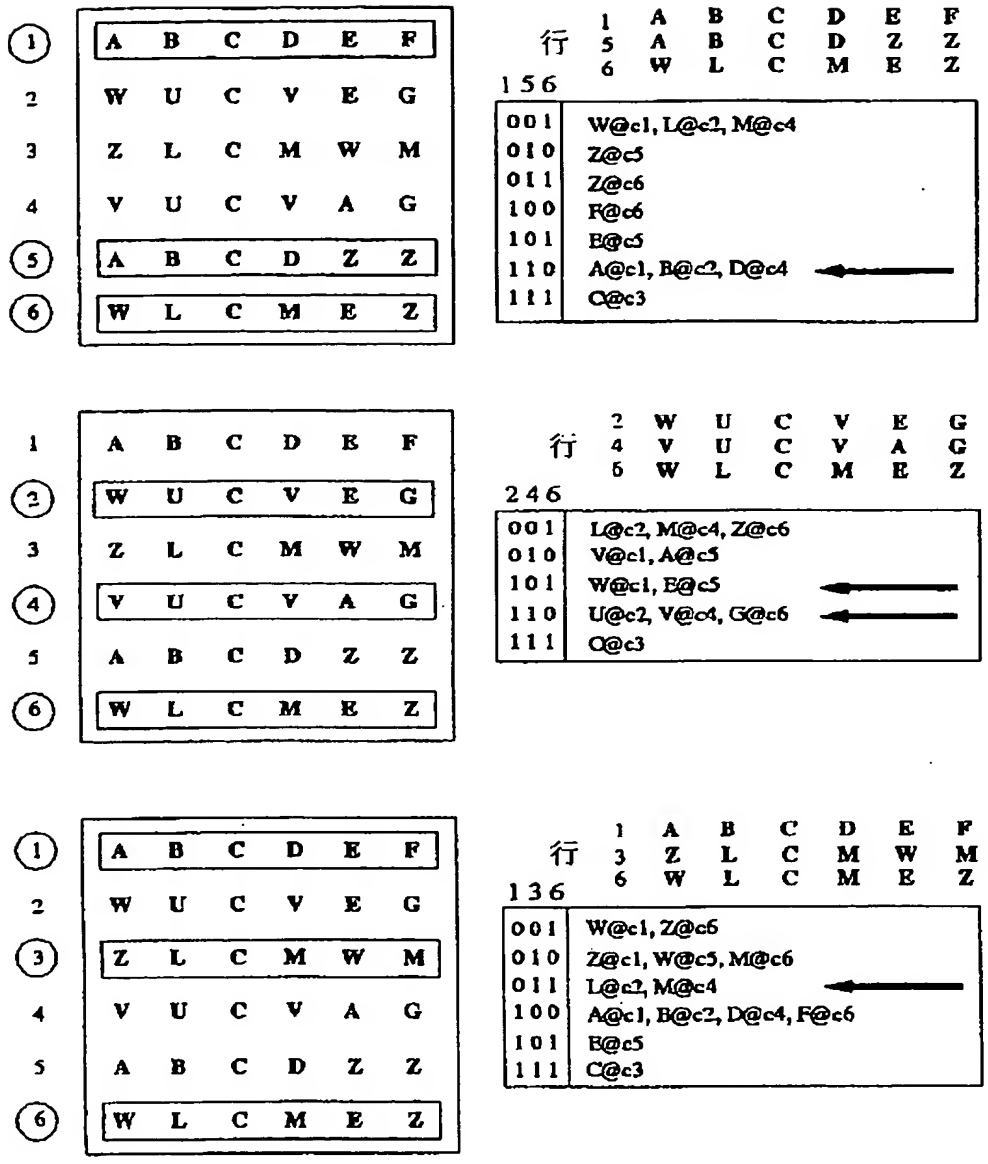

1 A B C D E F
2 W U C V E G
3 Z L C M W M
4 V U C V A G
5 A B C D Z Z
6 W L C M E Z
行 1 5 6
A B C D E F
A B C D Z Z
W L C M E Z
156
001 W@c1, L@c2, M@c4
010 Z@c5
011 Z@c6
100 F@c6
101 E@c5
110 A@c1, B@c2, D@c4
111 C@c3
1 A B C D E F
2 W U C V E G
3 Z L C M W M
4 V U C V A G
5 A B C D Z Z
6 W L C M E Z
行 2 4 6
W U C V E G
V U C V A G
W L C M E Z
246
001 L@c2, M@c4, Z@c6
010 V@c1, A@c5
101 W@c1, E@c5
110 U@c2, V@c4, G@c6
111 C@c3
1 A B C D E F
2 W U C V E G
3 Z L C M W M
4 V U C V A G
5 A B C D Z Z
6 W L C M E Z
行 1 3 6
A B C D E F
Z L C M W M
W L C M E Z
136
001 W@c1, Z@c6
010 Z@c1, W@c5, M@c6
011 L@c2, M@c4
100 A@c1, B@c2, D@c4, F@c6
101 E@c5
111 C@c3

図５

【図５】

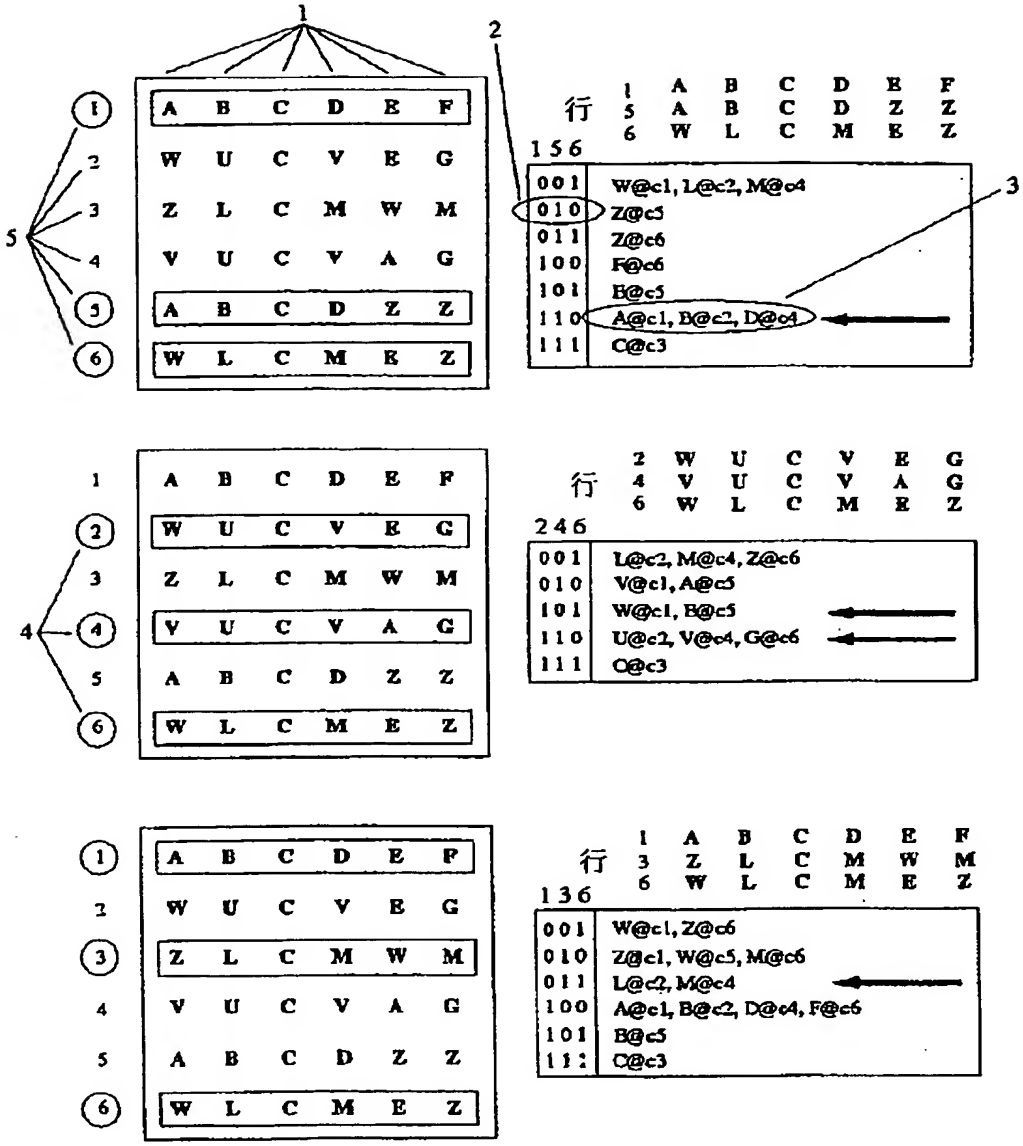

1
2
3
4
5
6
A B C D E F
W U C V E G
Z L C M W M
V U C V A G
A B C D Z Z
W L C M E Z
行 1 5 6
A B C D E F
A B C D Z Z
W L C M E Z
156
001 W@c1, L@c2, M@c4
010 Z@c5
011 Z@c6
100 F@c6
101 E@c5
110 A@c1, B@c2, D@c4
111 C@c3
行 2 4 6
W U C V E G
V U C V A G
W L C M E Z
246
001 L@c2, M@c4, Z@c6
010 V@c1, A@c5
101 W@c1, B@c5
110 U@c2, V@c4, G@c6
111 C@c3
行 1 3 6
A B C D E F
Z L C M W M
W L C M E Z
136
001 W@c1, Z@c6
010 Z@c1, W@c5, M@c6
011 L@c2, M@c4
100 A@c1, B@c2, D@c4, F@c6
101 B@c5
11: C@c3

図５ａ

【図６】

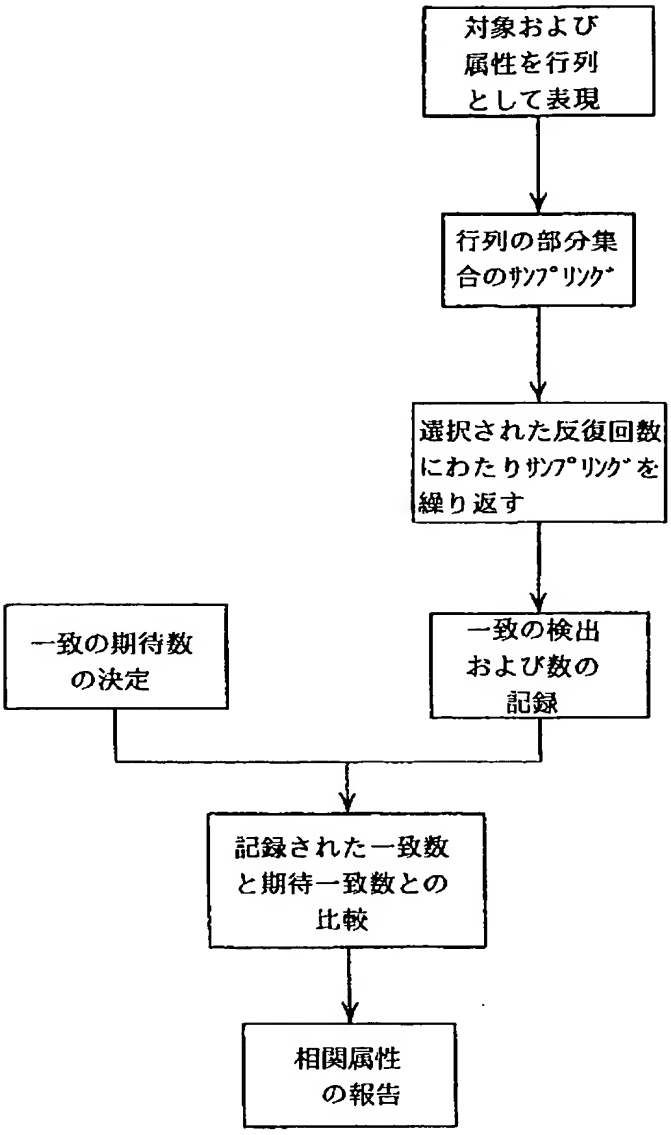

対象および属性を行列として表現
行列の部分集合のｻﾝﾌﾟﾘﾝｸﾞ
選択された反復回数にわたりｻﾝﾌﾟﾘﾝｸﾞを繰り返す
一致の期待数の決定
一致の検出および数の記録
記録された一致数と期待一致数との比較
相関属性の報告

図６

【図７】

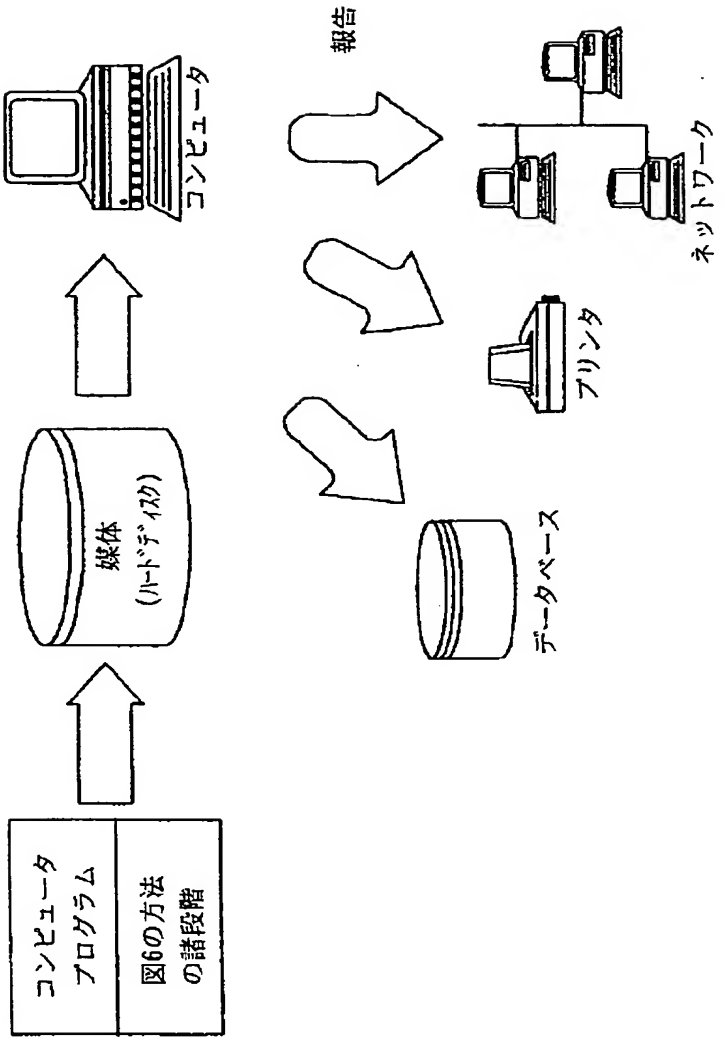

コンピュータ
プログラム
図6の方法
の諸段階
媒体
(ハードディスク)
コンピュータ
報告
ネットワーク
プリンタ
データベース

図７

【図８】

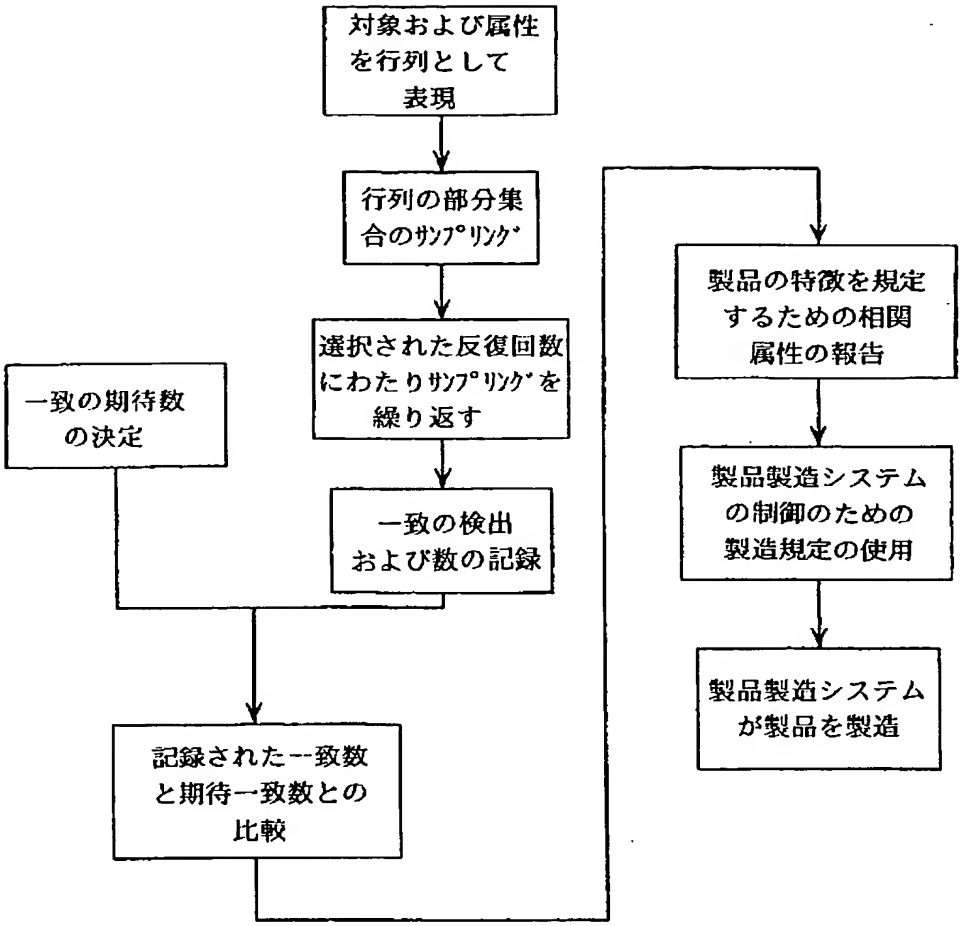

対象および属性を行列として表現
行列の部分集合のｻﾝﾌﾟﾘﾝｸﾞ
選択された反復回数にわたりｻﾝﾌﾟﾘﾝｸﾞを繰り返す
一致の検出および数の記録
一致の期待数の決定
記録された一致数と期待一致数との比較
製品の特徴を規定するための相関属性の報告
製品製造システムの制御のための製造規定の使用
製品製造システムが製品を製造

図８

【図９】

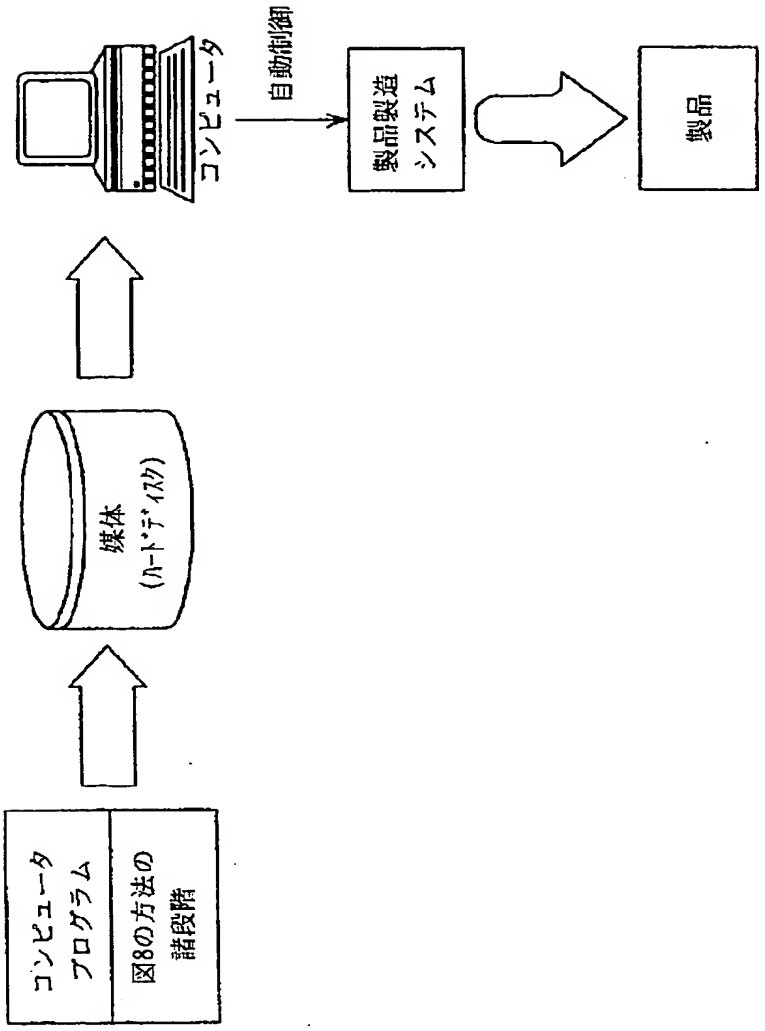


図９

【図１０】

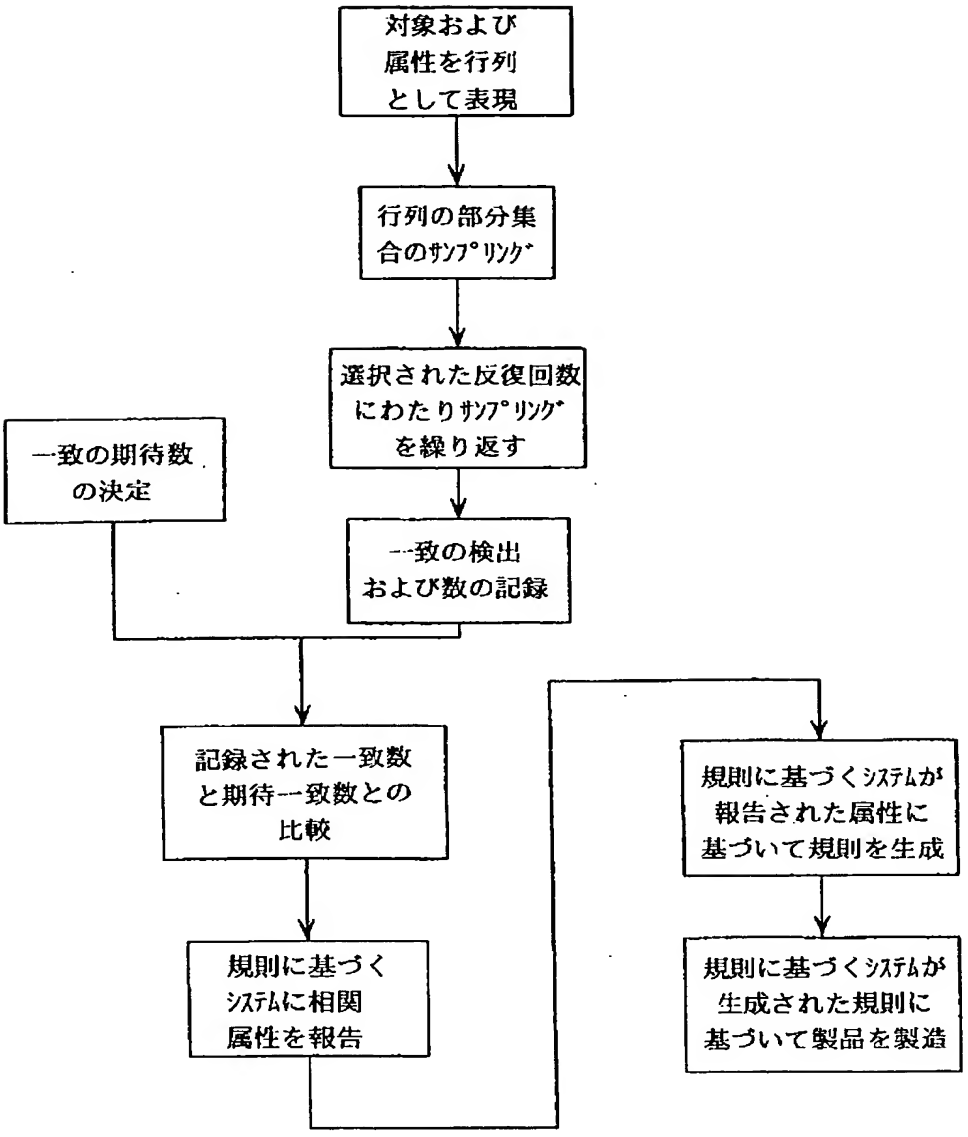

対象および属性を行列として表現
行列の部分集合のｻﾝﾌﾟﾘﾝｸﾞ
選択された反復回数にわたりｻﾝﾌﾟﾘﾝｸﾞを繰り返す
一致の期待数の決定
一致の検出および数の記録
記録された一致数と期待一致数との比較
規則に基づくｼｽﾃﾑに相関属性を報告
規則に基づくｼｽﾃﾑが報告された属性に基づいて規則を生成
規則に基づくｼｽﾃﾑが生成された規則に基づいて製品を製造

図１０

【図１１】

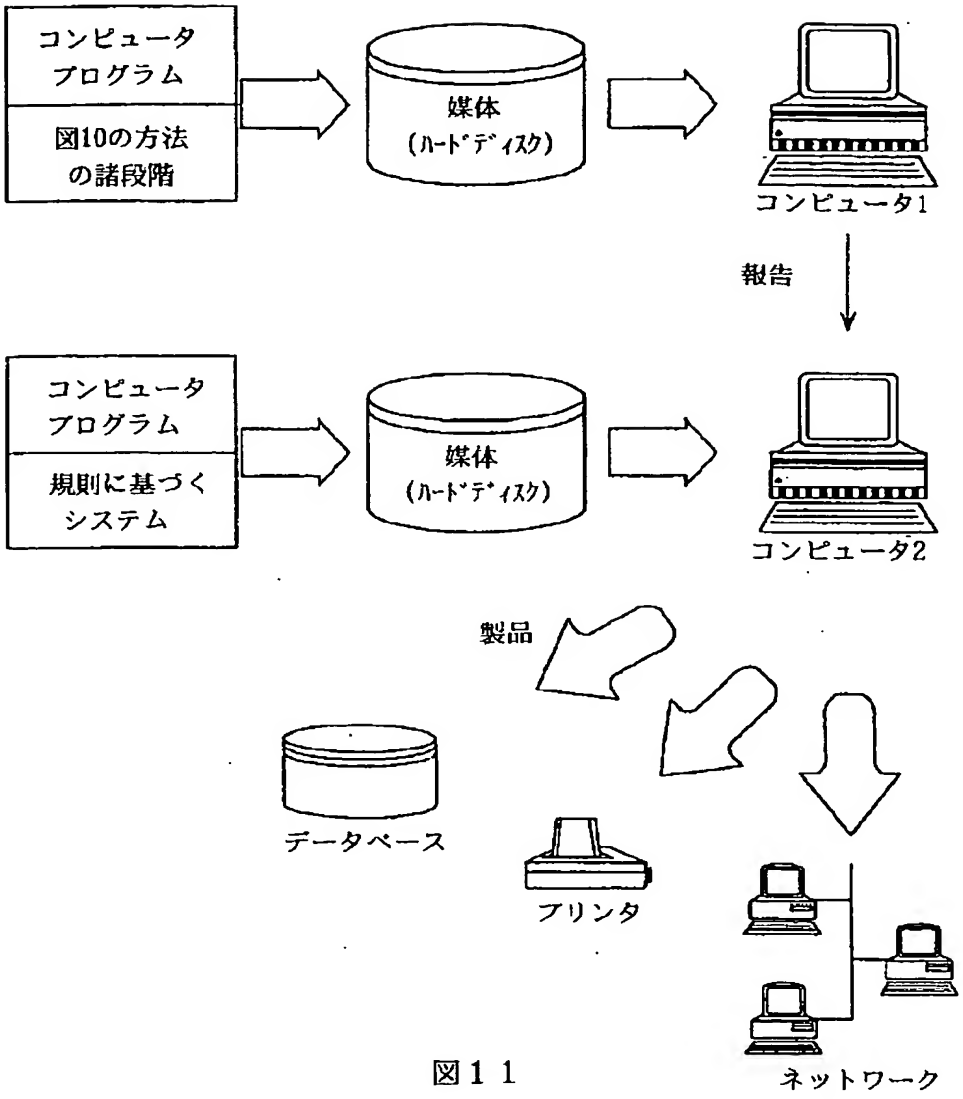

コンピュータ
プログラム
図10の方法
の諸段階
媒体
(ﾊｰﾄﾞﾃﾞｨｽｸ)
コンピュータ1
報告
コンピュータ
プログラム
規則に基づく
システム
媒体
(ﾊｰﾄﾞﾃﾞｨｽｸ)
コンピュータ2
製品
データベース
プリンタ
ネットワーク

図１１

【図１２】

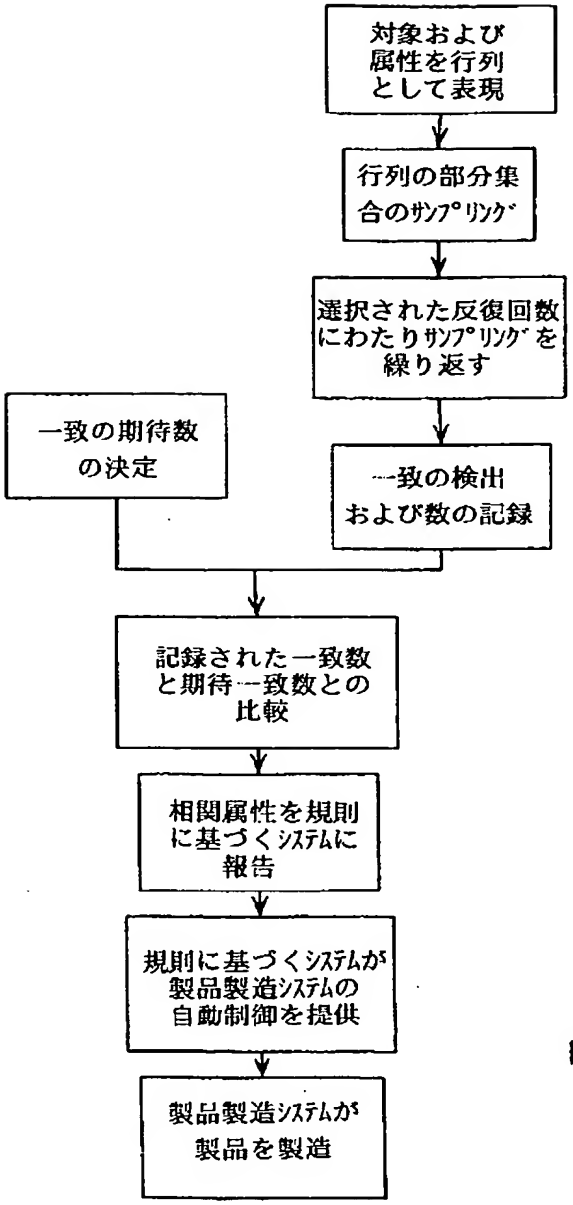

対象および属性を行列として表現
行列の部分集合のｻﾝﾌﾟﾘﾝｸﾞ
選択された反復回数にわたりｻﾝﾌﾟﾘﾝｸﾞを繰り返す
一致の期待数の決定
一致の検出および数の記録
記録された一致数と期待一致数との比較
相関属性を規則に基づくｼｽﾃﾑに報告
規則に基づくｼｽﾃﾑが製品製造ｼｽﾃﾑの自動制御を提供
製品製造ｼｽﾃﾑが製品を製造

図１２

【図１３】

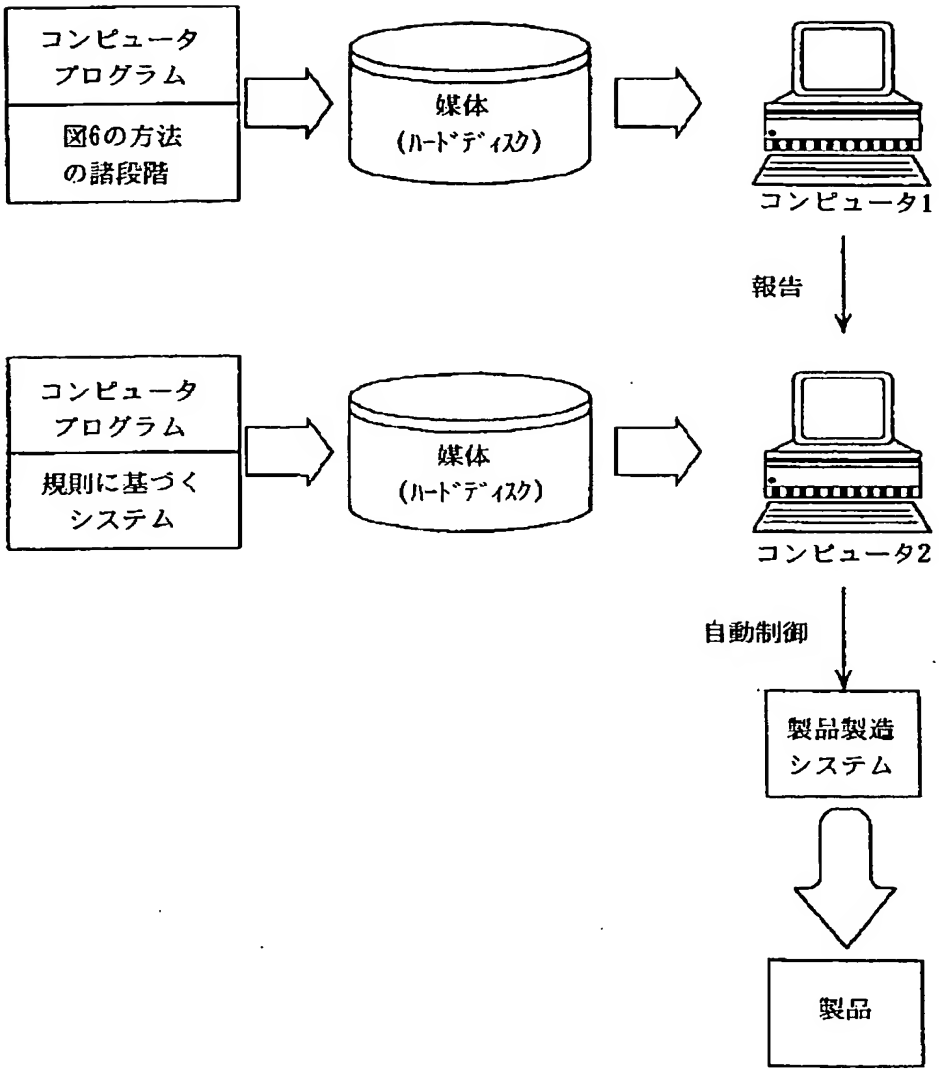

コンピュータ
プログラム
図6の方法
の諸段階
媒体
(ﾊｰﾄﾞﾃﾞｨｽｸ)
コンピュータ1
報告
コンピュータ
プログラム
規則に基づく
システム
媒体
(ﾊｰﾄﾞﾃﾞｨｽｸ)
コンピュータ2
自動制御
製品製造
システム
製品

図１３

【図１４】

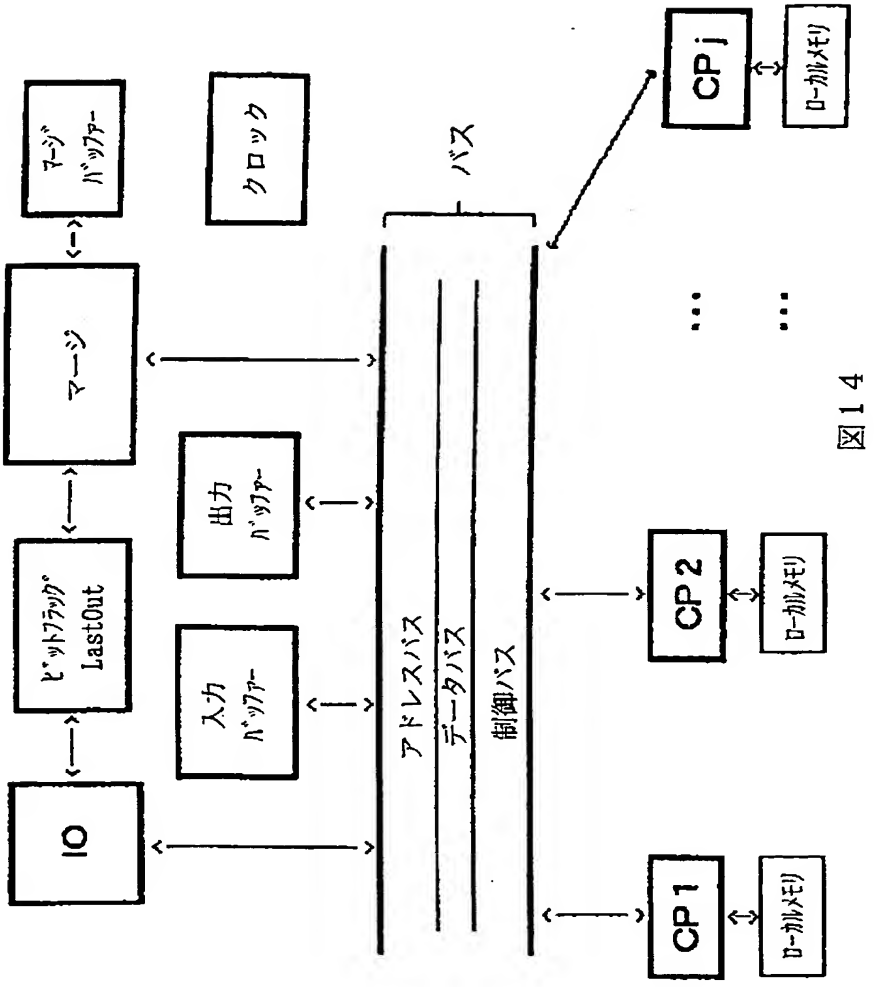

マージ バッファー
クロック
CPj
ローカルメモリ
バス
マージ
出力 バッファー
ビットフラッグ LastOut
入力 バッファー
IO
アドレスバス
データバス
制御バス
CP2
ローカルメモリ
CP1
ローカルメモリ

図１４

【図１５】

残基	1	2	•••	i	•••	j	•••
配列							
1	I	L		W		G	
2	S	C		G		W	
3	L	C	•••	Y	•••	A	•••
4	A	P		W		G	
5	S	A		Y		A	
6	R	R		G		Y	
•	•	•		•		•	
•	•	•		•		•	
•	•	•		•		•	
M-1	C	P		W		G	
M	L	I	•••	A	•••	Y	•••

整列化された
相同配列の
ファミリー

大きな側鎖⇒小さな側鎖

小さな側鎖⇒大きな側鎖

図１５

【図１５】

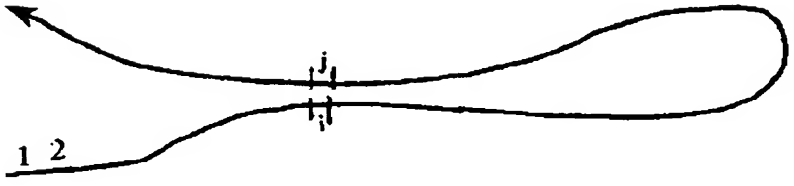
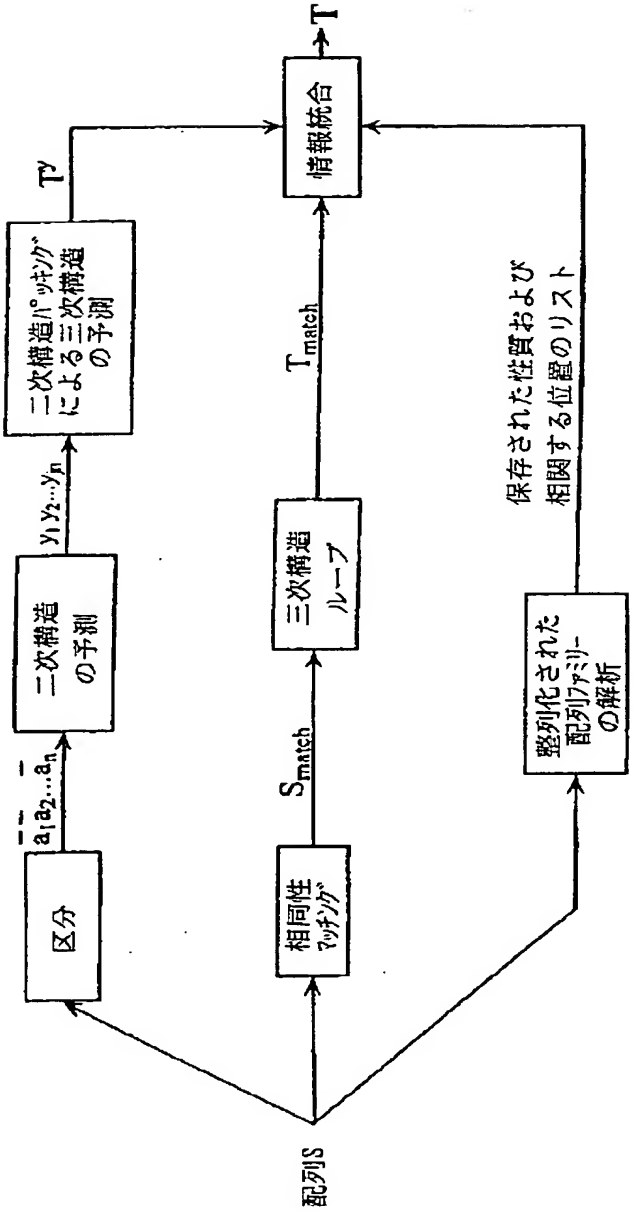


図１５Ａ

【図１６】

図１６


配列S
区分
相同性マッチング
整列化された配列ファミリーの解析
$a_1a_2...a_n$
二次構造の予測
$y_1y_2...y_n$
二次構造パッキングによる三次構造の予測
S_{match}
三次構造ループ
T_{match}
保存された性質および相関する位置のリスト
Ty
情報統合
T

【国際調査報告】

INTERNATIONAL SEARCH REPORT

tnt tional Application No
PCT/CA 98/00273

A. CLASSIFICATION OF SUBJECT MATTER
IPC 6 G06F17/30

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 6 G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	B. T. M. KORBER ET AL: "Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope proteins: An information theoretic analysis" PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES, vol. 90, - August 1993 US, pages 7176-7180, XP002069939 cited in the application see the whole document --- -/--	1-20,23, 24, 29-34, 37-54

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
6 July 1998	20/07/1998

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Abram, R

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

Int. Application No
PCT/CA 98/00273

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	A. F. W. COULSON ET AL: "Protein and nucleic acid sequence database searching: a suitable case for parallel processing" THE COMPUTER JOURNAL, vol. 30, no. 5, October 1987, CAMBRIDGE, GB, pages 420-424, XP000049651 see the whole document	1-20,23, 37-54
A	R. GUIGÓ ET AL: "Inferring Correlation between Database Queries: Analysis of Protein Sequence Patterns" IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, vol. 15, no. 10, October 1993, NEW YORK, NY, US, pages 1030-1041, XP000403522 see the whole document	1-20,23, 37-54
A	GB 2 283 840 A (FUJITSU LIMITED) 17 May 1995 see the whole document	1-20,23, 37-54
P,A	P. MICHAUD: "Clustering techniques" FUTURE GENERATION COMPUTER SYSTEMS, vol. 13, November 1997, NL, pages 135-147, XP004099490 see the whole document	1-20,23, 37-54
A	A. S. LAPEDES ET AL: "Use of Adaptive Networks to Define Highly Predictable Protein Secondary-Structure Classes" MACHINE LEARNING, vol. 21, no. 1/2, October 1995 - November 1995, BOSTON, MA, US, pages 103-124, XP002069940 see the whole document	1-20,23, 37-54

INTERNATIONAL SEARCH REPORT

Inte ional Application No
PCT/CA 98/00273

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
GB 2283840 A	17-05-1995	JP 7274965 A	24-10-1995
		US 5598350 A	28-01-1997

フロントページの続き

(51)Int.Cl.[7]	識別記号	FI	テーマコード(参考)
G01N 33/68		G06F 17/30	350A
G06F 17/30	350	A61K 37/02	

(81)指定国 EP(AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OA(BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG), AP(GH, GM, KE, LS, MW, SD, SZ, UG, ZW), EA(AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GE, GH, GM, GW, HU, ID, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZW